Kinds of Agency and the Role of Reflective Endorsement

Bucelli, Irene

Awarding institution:
King's College London

King's College London

KINDS OF AGENCY AND THE ROLE OF REFLECTIVE ENDORSEMENT

Irene Bucelli
Ph.D. in Philosophy

Abstract

Human beings act, and it is intuitive to think that they are agents in a rather unique way, one that is different from other animals. This intuition has led some philosophers to think that human agency exhibits the distinctive feature of being self-controlled, self-governed and autonomous. Some authors identify a form of agency, sometimes defined as full-blown, strong or *par excellence,* with which we can only credit human beings, and which is taken to be distinctive of some human actions. Within this framework, a prominent understanding of the notion of self-governance conceives it in terms of the agent directing and governing his own practical thought and actions. This position not only considers that self-governance is required for our behaviour to count as a full-blown action, but it also identifies the condition of self-governance with the agent's reflective endorsement: with the commitment to his own doings by means of his reflective capacities.

This thesis asks whether it makes sense to distinguish two kinds of agency, one of which is specifically human and expresses the agent's self-governance. I take issue with the prominent attempt to make sense of the distinction in terms of reflective endorsement and I claim that there are foundational reasons why accounts that employ this notion are unsatisfactory. In particular I argue that reflective endorsement approaches to agency are too restrictive and not realistic. While the main aim of the thesis is to criticize the fundamental assumptions that ground this highly prominent view of human actions, the difficulties that will emerge from my discussion will point at the desiderata for an alternative theory of agency, which will remain as a working hypothesis to develop in further work.

Contents

Introduction

The Question

Human beings act, and it is intuitive to think that they are agents in a rather unique way, one that is different from other animals. Human beings can act for reasons. Moreover, human actions can be subject to normative considerations, raising the issue of whether one should or should not act in a certain way under certain circumstances.

This intuition has led some philosophers to think that human agency exhibits the distinctive feature of being self-controlled, self-governed and autonomous. Some authors identify a form of agency, sometimes defined as full-blown, strong or *par excellence,* with which we can only credit human beings, and which is taken to be distinctive of some human actions. Within this framework, a prominent understanding of the notion of self-governance[1] conceives it in terms of the agent directing and governing his own practical thought and actions. This position not only considers that self-governance is required for our behaviour to count as a full-blown action, but it also identifies the condition of self-governance with the agent's reflective endorsement: with the commitment to his own doings by means of his reflective capacities.

Consider the following passages that illustrate this idea:

> I believe that human beings differ from the other animals in an important way. We are self-conscious in a particular way: we are conscious of the grounds on which we act, and therefore are in control of them. When you are aware that you are tempted, say, to do a certain action because you are experiencing a certain desire, you can step back from that connection and reflect on it. You can ask whether you should do that action because of that desire, or because of the features that make it desirable. And if you decide that you should not then you can refrain. This means that although there is a sense in which what a non-human animal does is up to her, the sense in which what you do is up to you is deeper.[2]

> When a person acts because she desires or intends or the like, we sometimes do not want to say simply that the pro attitude leads to the action. In those cases we suppose, further, that the agent is the source

[1] For important more or less recent examples see, for example, Frankfurt, 1971; Bratman, 2007; Korsgaard, 2009; Velleman, 2007a.
[2] Korsgaard, 2009, p. 19.

> of, determines, directs, governs the action, and it is not merely the locus of a series of happenings, of causal pushes and pulls. A skeptic might doubt that there really is an important distinction between (merely) motivated behaviour and action determined and governed by the agent and it is true that in any case of motivated behaviour the agent in some sense acts. Nevertheless many of the cases that suggest a gap between desire-based motivation that is and that is not appropriately related to the agent's normative deliberation also suggest a distinction between (merely) motivated behaviour and, as I will call it, full blown agency. An agent moved by desires of which he is unaware, or on which he is incapable of reflecting, or from whose role in action he is, sometimes say, estranged, seems himself less the source of the activity than a locus of forces.[3]

According to these views there are two kinds of agency: on the one hand we find merely purposive behaviour, which some non-human animals can also perform; on the other hand there are full-blown actions, which are specifically human and spring from the agent's self-governance and from his control over his motives. Furthermore, these views stress the role played in human agency by the specific capacity for self-consciousness, through which one comes to reflectively endorse one's motives and perform actions that are self-governed. This means that while animals are determined by their urges, human beings are capable of being *self*-determined: we can control our behaviour so that, by committing to some of our motives, we regard them as reasons and perform the relevant behaviour in light of them. Thus, because of our cognitive reflective capacities, our actions are not mere outcomes of whatever state we are in. Moreover, it is important to notice that the human specific conscious reflective capacities are taken to give a clear threshold between these kinds of agency.

This thesis asks whether it makes sense to distinguish two kinds of agency, one of which is specifically human and expresses the agent's self-governance. I take issue with the prominent attempt to make sense of the distinction in terms of reflective endorsement and I claim that there are foundational reasons why accounts that employ this notion are unsatisfactory. In particular I argue that reflective endorsement approaches to agency are too restrictive and not realistic. While the main aim of the thesis is to debunk the fundamental

[3] Bratman, 2007, p. 91.

assumptions that ground this highly prominent view of human actions, the difficulties that will emerge from my discussion will point at the desiderata for an alternative theory of agency, which will remain as a working hypothesis to develop in further work.

Methodological Considerations

I will focus my discussion on two philosophers that are central figures in the framework of those who understand self-governance in terms of reflective endorsement: Christine Korsgaard and David Velleman. The reason for this choice is that both Korsgaard and Velleman are proponents of theories of agency that are, at the same time, economic and ambitious. They are economic because they share the idea that there is *one constitutive feature* that we must uncover to characterize self-governed agency. And they are ambitious because they consider that our understanding of what actions are is all we need in order to give an account of the nature of normativity or practical reason. Because of these characteristics, the discussion of their models is particularly compelling within the reflective endorsement framework. Reflective endorsement theories of agency, as I have briefly sketched them above, make two fundamental claims: firstly, they identify self-governance, which they conceive as involving one's reflective capacities and is thus linked to reflective endorsement, as defining a kind of agency that is specifically human and that is different from the agency we can credit animals with. Secondly, they consider that this distinction between kinds of agency is relevant for issues of moral responsibility and answerability, inasmuch as it is by employing one's self-governing reflective capacities that one counts as acting for reasons. While their models give rather different accounts of reflective endorsement, Korsgaard and Velleman present similarities in the way they connect these two claims. Their "constitutivist"[4] approaches to agency see the descriptive task of understanding the nature of actions and the conditions of agency as bearing directly on and, in fact, providing the grounding for, normative debates.

[4] Here I employ the same terminology used by Enoch, 2006.

Talk of a "constitutivist approach" here refers to the fact that both theories strive to establish the *constitutive* conditions of actions and, as we shall see, these conditions ascribe a central role to the agent's reflective endorsement. For example, in Korsgaard, we find the following thought: in order to know what it takes for a house to be a good house, we need to understand what houses are and what their constitutive function is. This gives us standards to evaluate the extent to which a certain house fulfils this constitutive function and houses that do not meet these standards are to be considered defective, to the point that they can fail to be houses at all. Similarly, in relation to agency, the thought is that to understand the normative standards for actions, we first need an understanding of what actions are, what their constitutive function is. Actions that do not fulfil these constitutive standards are defective, and in fact they do not count as actions at all.

There is a second methodological limitation that I am going to set to my analysis. I just stressed how reflective endorsement approaches attempt to account for the specificity of human agency. I also highlighted how they argue for a close connection between this descriptive claim and normative issues. In what follows I will focus on the nature of agency and the plausibility of the description of it these views offer, and I will not dwell on the implications that these theories have for debates in moral philosophy. This will mean, for example, that I won't discuss the plausibility of Korsgaard's normative ethics.

Furthermore, I will avoid assessing related debates that seem to have important connections with these views of agency. For example, I will mention but will not discuss in detail the conception of rationality that underlies these two models, nor will I examine the commitments of these views to internalist or externalist conceptions of reasons for actions.[5] What I am going to focus on is rather the conception of reflective endorsement that these views define as constitutive of agency.

[5] See ibid. for a discussion of Velleman "quasi-externalism" see Velleman, 2000.

In order to have a clear characterization of what reflective endorsement amounts to, I believe that any account of agency that assigns to it a central role needs to answer three fundamental questions:

1) What is the relation between self-consciousness and self-governance and how do they relate to the actual production of behaviour? This question focuses on the way in which we should characterize motivation and it requires to a statement of the relation between reasons and actions.

2) What is the "self" that the expression "self-governance" refers to? Is it the same "self" as in "self-consciousness"?

3) What does the reflection involved in reflective endorsement amount to?

This question asks for a characterization of the psychological processes and the experience associated with reflective endorsement.

I will present Korsgaard's and Velleman's models and I will examine in detail how they answer these three questions. I will then consider whether or not these answers are satisfactory and my conclusion will be that they are not.

The Structure of this work

In chapter 1, I discuss Korsgaard's model and focus my critical analysis on her understanding of reflection in terms of deliberation. I show why a model of agency like hers, that assumes that actions necessarily require deliberation, is not satisfactory: it faces charges of producing a vicious regress and has the implausible consequence that very few of the things we do could actually count as actions. I consider what I think would be a distinctively Kantian solution that would allow Korsgaard to avoid this problem but I conclude that this solution is not available to her.

In chapter 2, I present Velleman's model and I show how his understanding of the reflective endorsement test in terms of narrative coherence seems to present advantages over Korsgaard's. Despite this I believe that Velleman's model also is too restrictive and is committed to implausible demands on our self-understanding. Moreover, I claim that the stress Velleman places on the

role played in actions by one's narrative self does not fit well with other parts of Velleman's model of agency. This leads me to discuss the two kinds of reflection that the model assumes and to show that they do not integrate with the metaphor of narration.

While chapters 1 and 2 focus on Korsgaard and Velleman respectively, chapters 3 and 4 examine the notion of self-consciousness and conscious control, which, I claim, generate the difficulties I highlighted in the first two chapters.

Chapter 3 aims at examining the notion of self-consciousness that emerged as crucial for both Korsgaard and Velleman. My starting point is a more detailed discussion of the idea of a 'narrative self'. This conception of the self has proved very influential both in philosophy and psychology and I devote a consistent part of the chapter to a detailed examination of it. Starting from Daniel Dennett's original formulation, I explore the metaphysical commitments that this notion has and how Velleman redefines it. I claim that the 'narrative self' should be understood as a narrative self-representation: an element of the subject's psychology which is neutral about the metaphysics of the subject. The attempt of this chapter is to show how both Korsgaard and Velleman share a conception of self-consciousness which is ultimately a fundamental problem for both theories because, I will argue, mischaracterizes the self-consciousness associated with human agency and is thus bound to generate some of the difficulties I highlight in chapters 1 and 2.

Finally, in chapter 4, I focus on the notion of conscious control which is employed by both Korsgaard and Velleman. I discuss it first in relation to the dimension of agentive awareness associated with controlled actions and, secondly, in relation to the psychological structure that such control requires for agency. I will argue that neither aspect finds a satisfactory treatment in these models. This unsatisfactory notion of control grounds the difficulties that I discuss in the first two chapters: it is the source of Korsgaard's problems in integrating aspects of the agent's experience with the psychological structure she requires for agency; and can be seen as underlying Velleman's ambiguities in defining the role of narrativity.

This overall discussion will lead me to conclude that Korsgaard's and Velleman's way of conceiving agency as distinguished in two kinds, one of which expresses the agent's self-governance through reflective endorsement, is fundamentally flawed. This negative conclusion leads me to point at some desiderata for a theory of agency that seeks to be realistic and still capture the specificity of human behaviour.

Chapter 1

Korsgaard's Constitutional Model of Agency

When we are acting, Christine Korsgaard claims, "it is as if there is something over and above your incentives, something which is *you*, and which chooses which incentives to act on".[1] This is the very phenomenon that Korgsaard's work aims at explaining: our experience, as agents, of being something over and above our various (possibly conflicting) desires and instincts. In her view, this phenomenological experience mirrors a metaphysical difference between actions and mere movements an agent can perform. In showing the conditions for the kind of agency that we, as humans, can enjoy, Korsgaard aims at showing how, when a person performs an action, this is not caused by some aspect or part of her, but is actually expressive of the agent as a whole so that the person is the author of her action, rather than caused to act by her own desires and inclinations. Korsgaard conceives of action as self-governed behaviour that can be performed exclusively by creatures that are capable of self-control and have the capacity for self-consciousness. Korsgaard is thus among those philosophers whose views find in the agent's reflective endorsement the necessary and sufficient condition for agency. In my discussion of her view I will start by presenting how she answers the three questions I have stressed any such theory needs to address, namely: (§1) what is the relation between self-consciousness and self-governance? (§2) What is the "self" of self-governance? (§3) What does reflection amount to?

[1] Korsgaard, 2009, *72*, p. 126.

1.1 What is the relation between self-consciousness and self-governance?

In answering this question on Korsgaard's behalf we firstly need to be clear about how self-consciousness figures in her story and what she takes self-governance to be.

Korsgaard wants to allow for the fact that animals are able to act, but animal agency is rather different from the form of agency that is particular to humans. In her view, an animal perceives a danger and its natural reactions are to run away, or to fight the intruder: in any case, its instincts are to undertake a certain course of action, and that is the action that it consequently performs. The animal's instinct determines the animal's action. So, while animals are conscious, inasmuch as they are capable of enjoying representations of the environment and their behaviour is responsive to such conscious representation[2], animals are not self-conscious. Self-consciousness is precisely what creates the psychic complexity[3] that gives human agency what Korsgaard believes are different constitutive standards and different conditions of possibility. Self-consciousness allows human beings the peculiar possibility of having a 'reflective distance' from their own mental states. Once I have distanced myself from my mental states, they cannot move me to act as they did in the case of lower forms of agency. The mere desire to run away from peril does not automatically move me to act: it is a merely potential ground for

[2] Korsgaard specifically describes animal action as being responsive to a conception of the environment and being a movement with intentional content. For animal actions to have an intentional content means that they have a certain purpose, in the light of which they can succeed or fail; rather than attributing some thought process to the animal. See ibid., pp. 96-97.

[3] Ibid., p. 125, 213; and Korsgaard, 1996, p.94.

me to act and I see it as such, and this means that I need a reason to act upon it. This also modifies the connection between mental state and action: it ceases to be one of causation and becomes one of motivation. So, faced with my instinct to run away, I can judge that this course of action really is the best and I end up running away. However, it could have been the case that, despite my desire to run away, I had reasons I consider good enough not to do so if, for example, in running away I would be abandoning a friend in a dangerous situation, or I would be proving myself a coward in front of the person I have a crush on; thus I decide on a different course of action. In endorsing a certain course of action, I am motivated by the awareness that my motives are good ones.

From these considerations one might draw the claim that, since I am conscious, I can be in control of my desires. This claim would seem false: my being conscious of a certain desire, one that I do not welcome and I would rather not have, might not enable me to control its impact on my behaviour. However, this is not what Korsgaard claims because according to her, self-consciousness is necessary, yet not sufficient, for action. Self-consciousness is a condition for self-governance, but for something to count as an action, and hence for my behaviour to count as self-governed, it needs to pass the reflective scrutiny at the end of which the agent commits to a certain course of action, avowing it for reasons. An action is self-governed iff it has passed the Reflective Endorsement test, and Korsgaard characterizes such a test in terms of the Kantian categorical imperative. Your maxim, once formulated, embodies your proposed reason. You then test it by the categorical imperative, that is,

you ask whether you can will it to be a universal law, in order to see whether it really is a reason. Universalizability is a condition on the form of a reason, and if a consideration does not meet this condition, then it is not merely outweighed—rather, it is not a reason at all.[4]

This means two things. First, it follows that in practical reasoning you formulate your action in terms of its means-end structure (in Kantian terms, you incorporate it within a maxim, which exhibits the means-end structure): so you will not just be 'visiting your parents' or 'helping your friends', you will be 'visiting your parents in order to make them happy' or 'helping friends, so that they owe you a favour'. And second, it means that in order to establish whether or not you have a reason to act according to your maxim you need to check whether your maxim can be universalized (or, in Kantian terms, can "count as a law"). The categorical imperative serves as a test inasmuch as it is a way to establish whether that particular action, with all its relevant features specified, can be universalised. If it can, one's motive counts as a reason.[5] So, for example, your helping friends so that they will owe you a favour, and you will ultimately be able to take advantage of them, will not make a maxim that can pass the universalizability test. If, however, you help your friends out of your genuine interest in their well-being, your maxim will pass.[6]

[4] Korsgaard, 2009, p. 51.
[5] This means testing the action's form. Korsgaard has an argument to prove that reasons are universal, which will not be discussed here. Ibid., pp. 71-76.
[6] A qualification to this consideration is that for really passing the universalizability test you must not make your friends an exception, so that you break the rules of morality in order to help them or you impact someone else's autonomy and well-being by doing so etc.

To sum up, only through endorsement can I achieve that psychic unity that my being a self-conscious being sets as a condition for me to aim at acting at all. This is so because in experiencing the psychic complexity produced by my self-conscious capacities, I recognize the need for reasons in order to undertake a certain course of action. And this means that, on the one hand, in humans, one's original motivation to act coincides with one's motivation to act for reasons. On the other hand, it means that only by applying the categorical imperative test (which establishes whether my motives count as reasons and my maxims count as laws), can I fulfil the requirements of agency. Self-consciousness obliges me to act in accordance with self-governance if I am to act at all.

According to Korsgaard's model, then, in acting I determine the "kind of causality"[7] I am as an agent: I act under the idea that I am self-governed, that my action is attributable to the whole person, not to something *within* me. This means that while I will to act, I necessarily also will the principles that are constitutive of agency. Korsgaard claims that this makes "autonomy" a metaphysical property[8] of action, that sets the normative standards for behaviour to count as an instance of agency. This might seem hopelessly vague, but it can be given further specification once we have shown that in determining the "kind of causality" I am, I am determining "who" I am. This is what I will examine in the next section.

[7] Ibid., pp. 82-83, 127.
[8] Korsgaard, 1999, p. 13.

1.2. What is the "self" of self-governance?

Korsgaard claims that a self-governed action is an action governed by a self-conception. As we shall see, the categorical imperative will play a role in determining the self-conception that can guide our actions. A preliminary consideration to make before considering how this is so is what Korsgaard takes this 'self' that governs action through one's self conception to be.

Korsgaard defines one's self-conception as a kind of practical identity. The notion of practical identity is a part of Korsgaard's philosophy that has been widely discussed[9] and appeals to the intuitive idea that each of us recognizes different values and concerns that are associated with various social and personal roles that we have in our lives. Korsgaard believes that your identity as a person refers to a plurality of these roles you value and consider relevant in your life, so that in defining who you are you put these many roles together. We can consider ourselves as belonging to a certain ethnic or social group, or as being in a certain relationship with others (I am someone's daughter, or friend, I am a student or practise a certain job or sport, I belong to some political party or religion etc.). These practical identities are not mere facts[10] about individuals but rather they are descriptions under which one values oneself and involve roles one endorses. This means that, for example, although

[9] See, for example, Cohen, 1996, or Velleman, 2006, pp.284-311.

[10] Another way in which our practical identities are not to be considered as some facts is that these social roles are not set in stone, independently from my understanding of them: what being a good friend amounts to, or a good daughter, or a good Christian, and the obligations they entail, the values they promote, is open to my interpretation of those roles. I can, for example, consider myself a Catholic but still see no obligation to go to church every Sunday. This also means that practical identities are open to originality: "I can find new ways to be a friend" (Korsgaard, 2009, p. 101). Therefore, they are open to change, since my idea of what is important about them can vary over time.

I might be born and raised in a catholic family, if I do not consider myself a Catholic, then this is not a practical identity of mine. While our particular group memberships and relationships are, in one sense, a matter of chance in our lives (I just happened to be born in a certain time and place etc.), in order to be part of my self-conception, these contingencies are required to figure among the descriptions under which I value myself. The person's identity is the unity of these many practical identities, as a unified description of oneself and 'it is not that you have a personal identity you might or might not be conscious of. If you are not conscious of it, than you don't have it'.[11]

The most important characteristic of this conception of the self is that it must in itself be considered as an activity of unification of these different practical identities. Given the psychic complexity that, we have seen, is the main consequence of human self-consciousness, this activity is necessary to gain psychological unity. Because it is an activity, my self-conception is something I am in charge of: it is the result of my commitment and as such embodies the principles that ultimately guide my action. Since the categorical imperative is the constitutive principle of agency, my very identity as a moral agent is the one that guides action. Practical identities are sources of reasons for actions, and they also embody the guiding principles of choice that ultimately are the agent's specific contribution to the action.[12]

[11] Korsgaard, 2008, p.214n.

[12] Korsgaard discusses extensively how self-governed action requires a self-conception of a particular kind: since the categorical imperative is the constitutive principle of agency, actions that result from alternative principles are less than self-governed. See 2009, pp.161-171

Our practical identities can fail to guide us in the way required for the resulting behaviour to count as autonomous, which, for Korsgaard, means that it does not count as an actual instance of action. This is the case when, for example, a fanatic is committed to act upon the unreflected values of his sectarian ideology. The behaviour that such an identity dictates is not something that the person endorses for reasons, since the maxims such an ideology commits him to are not truly legitimately universalizable. In acting out such a practical identity then the agent is not truly autonomous, but is rather "enslaved by his dominating obsession".[13]

Before looking at the way in which Korsgaard answers our third question and characterizes the kind of reflection at work in reflective endorsement, I want to stress how, according to her, in acting the agent has a conception of what she is doing and why, which means that she has a conception of the principle on the basis of which she chooses, but this also coincides with a conception of her practical identity. In fact, Korsgaard adds, this is the kind of self that an animal cannot have, even if it is conscious and has an egocentric framework through which it interacts with the environment. Animals are not self-conscious and do not have a conception of themselves as occupying such an egocentric perspective. We can then see how kinds of selfhood and kinds agency mirror one another, since they derive from the same, peculiarly human, capacity for self-consciousness.

Korsgaard's model of agency allows for degrees of agency,[14] and allows for degrees of identity. Degrees of agency can be conceived as a spectrum with, at

[13] Korsgaard discusses such cases as ones of "tyrannical souls". See Korsgaard, 2009, p. 171.
[14] Ibid., 174.

one end, someone like a psychopath or a maniac, in whose actions the boundary between animal agency and human agency is blurred; and at the other end, we find the fully autonomous agent who is aware of the principles on which he acts. Degrees of identity therefore imply degrees of unification. As we have seen, being fully yourself corresponds to being fully autonomous. This means that there are degrees of unification inasmuch as something other than you – than your judgement and reflective activity – is involved in your choices. The presence of an alternative force, which is alien to the principle of practical reasoning opens the possibility for your unity to fall apart, since there can be an occasion in which not *you*, but, instead, this force – a desire or an inclination for example – causes your action, which will not then be a full action.[15] A paradigmatic example of this would be a wanton who is completely at the mercy of his impulses: he does not actively reflect on his motives and he passively follows the lead of his desires. He cannot be said to behave for reasons, if acting for reason involves a form of reflective success. And equally, his identity is scattered among his different competing inclinations.

I will conclude this section by stressing something that will turn up again, with important consequences, in the following discussion: the fact that Korsgaard conceives of reflective endorsement ultimately as a form of "identification". This is an idea that we find in other theories of reflective endorsement, prominently in Harry Frankfurt, but Korsgaard's understanding of identification is different from Frankfurt's because, for her, this process is not merely volitional: it is the result of the application of a reflective test, the

[15] Korsgaard draws a parallel between her model and Plato's *Republic*. For a full discussion of Plato's different constitutions and the kind of souls associated with each one cf. Korsgaard, 1999, pp. 17-20 and, 2009, pp. 135-152.

categorical imperative. My endorsing an action is my identifying with the principle it expresses, which in turn is my ascribing to myself a certain practical identity. If I revise my judgement, and abandon my endorsement, I am no longer able to think of myself under that particular description. The evaluative component associated with the categorical imperative is at the core of Korsgaard's Constitutivism, whose key ideas I briefly sketched in my introduction: this means that Korsgaard can deny that agents who wholeheartedly endorse principles other than the categorical imperative are genuine agents. One cannot wholeheartedly be endorsing the principle of self-interest or some compulsive belief or desire, as in the case of the sectarian fanatic: when someone is truly acting, he is committed to the categorical imperative all along, since he behaves under the idea of being an agent. The volitional unity of agents who fail to do this, therefore, is merely apparent.

My presentation of Korsgaard's model of agency is almost complete. In this section I showed how Korsgaard thinks of action as governed by a self-conception, which is a kind of practical identity. In acting, the agent has a conception of the principle on the basis of which she chooses, but this also coincides with a conception of her practical identity. In the previous section I showed how Korsgaard links self-consciousness to self-governance by ascribing a constitutive role to the categorical imperative: this principle is the fundamental test for the endorsement that is necessary for my motives to count as reason. This means that while I will to act, I necessarily also will the principles that are constitutive of agency. These claims give us an understanding of endorsement as a form of identification. The last step in

giving a full account of Korsgaard's position requires us to understand what it means for this endorsement to be reflective.

1.3. What does reflection amount to?

It seems clear that any theory that relies on the notion of reflective endorsement needs to provide us with some understanding of the process underlying the endorsement and what it is like for the agent to endorse something. In numerous passages of her work Korsgaard describes the reasoning associated with reflective endorsement in terms of a deliberative process.

> I perceive, and I find myself with a powerful impulse to believe. But I back up and bring that impulse into view and then I have a certain distance. Now the impulse does not dominate me and now I have a problem. Shall I believe? Is this perception really a reason to believe? I desire and I find myself with a powerful impulse to act. But I back up and bring that impulse into view and then I have a certain distance. Now the impulse does not dominate me and now I have a problem. Shall I act? Is this desire really a reason to act? The reflective mind cannot settle for perception and desire, not just as such. It needs a reason. Otherwise, as long as it reflects, it cannot commit itself or go forward.[16]

> *When you deliberate,* it is as if there is something over and above your incentives, something which is *you*, and which chooses which incentives to act on.[17]

> Kant offers us what I think of as a "testing" rather than a "weighing" model of reasons. On his view, the way you are supposed to deliberate is to formulate a maxim, stating the complete package of considerations that together favour the performance of a certain action. Of course, some of the reasoning that I described above, the marshalling of the relevant considerations, will still go on, but now it will be part of the work of formulating the maxim. […] Your maxim, once formulated, embodies your proposed reason. You then test it by the categorical imperative, that

[16] Korsgaard, 1996, p.93.
[17] See ibid. p. 100, and Korsgaard, 2009, pp. 72, 126.

> is, you ask whether you can will it to be a universal law, in order to see whether it really is a reason.[18]
>
> If the problem is that our perceptions might not withstand reflective scrutiny, then the solution is that they might. We need reasons because our impulses must be able to withstand reflective scrutiny. We have reasons if they do. The normative word 'reason' refers to a kind of reflective success.[19]

So we see that for Korsgaard an agent is motivated by his own recognition of the appropriate conceptual connection between his beliefs and desires and needs a reflective process to test his own motives and take them as reasons. In this sense, acting for reasons requires a form of reflective success and is determined by the categorical imperative test.

I will discuss in detail the notion of deliberation in the next section, since I believe that it causes serious problems for Korsgaard's picture of agency. For now, it is enough to say that Korsgaard sees as paradigmatic cases of agency ones in which the agent performs the relevant behaviour (like cancelling your appointment with the dentist) in light of motives that count as reasons, because they have passed the reflective endorsement test (so, cancelling the meeting because there has been an opening for the operation you were long awaiting would count as a good reason to act, while cancelling the meeting because you are scared of dentists would not).[20] Moreover, it is important to notice that one could deliberate while failing to adopt the principles constitutive of agency (one could adopt principles other than the categorical imperative or adopt no principle at all). In light of this we can understand how the fact that one's behaviour is the result of practical deliberation is necessary for action but deliberation is not sufficient for agency: only when you

[18] Korsgaard, 2009, p. 51.
[19] Korsgaard,1996, p. 93.
[20] Something I will not discuss here is the fact that Korsgaard considers reasons as "provisionally universal", meaning that for Korsgaard our maxims must take the form of a universal law (which is to say that they must conform to the categorical imperative) and willing them 'universally' means that we commit ourselves to act "unless there is a good reason why not" (2009; p.73). This clause allows her to specify that there are background conditions in everyday life that can be accounted for by the model. It makes it possible for Korsgaard to be less demanding than it appears in requiring the universalizability test necessarily: cancelling a meeting with a friend because I don't feel like it might (surprisingly if we were to take universalizability as absolute) pass the test, once we have noticed that the meeting was not important, that the cancellation does not produce any particularly bad results for the other party and my not feeling like it is the consequence of severe stress conditions due to work etc.

deliberate in the light of the rational principles constitutive of agency, which involve the categorical imperative, do you achieve the volitional unity that belongs to full agency.

1.4 Problems with Korsgaard's view

I shall argue that, through the constitutive conditions for agency identified in this model, Korsgaard restricts severely the range of full actions, with the result that her model gives a counter-intuitive picture of agency. I will argue that this problem is due to her commitment to conceiving of reflection as deliberation and to her considering the categorical imperative as the ultimate test for reflective endorsement.

In order to argue for these claims I will consider in more depth the notion of deliberation at play in her work. Nowhere[21] does Korsgaard explain what she means by the term, but I will argue that the way we characterize deliberation has important bearing on her theory. In fact, I will show that different conceptions of deliberation can make Korsgaard's view either thoroughly implausible and too restrictive in limiting the range of actions (2.1); or very vague and hence in need of further specification (2.2).

I will highlight some very general characteristics of deliberation. Deliberation is a form of deliberate or voluntary activity, one that aims at a resolution about what to believe, in the case of theoretical deliberation; or about what to do, in the case of practical deliberation. All these aspects are really important when we consider how deliberation differs from other forms of thinking. In its aiming at a resolution, deliberation is different from, for example, fantasising or recalling, although these are also mental activities, that, like deliberation, can comprise many steps that I rehearse in my thought. In fact, these are different mental activities that one can employ when deliberating: e.g. in

[21] I am here considering primarily Korsgaard, 2009, but also 2008 and 1996.

deliberating about what to give to my mum for Christmas I can recall past years, past presents and her reactions etc.. However, deliberation is also different from other ways of determining what to do: it is different from deciding that does not involve any conscious reasoning. Different from when a thought strikes you that you left the tub running or that you have forgotten to take your medicines. Also, it is a deliberate activity with a different phenomenology: there is something different in the subject's experience of what it is like to try and get to a resolution C and having that same thought C just dawning on you or occurring unintentionally. This is not to say that in deliberating you have an intention in advance to have a particular thought: you do not intend to resolve to buy your mum a certain sweater. But you do have an intention to think of a thought that holds particular relations of inference to others you are having[22] and you have the intention to get to a conclusion about what would be best to buy, whatever this turns out to be.

Moreover, it is important to notice that reason-giving is a practice that produces considerations much like the ones that we can imagine an agent to entertain during deliberation. Nevertheless, the similarity between the practice of reason-giving and deliberation should not obscure the fact that one can act or believe for reasons without it involving deliberation. To have a belief, a desire or an intention that we have reasons for is to be committed to something: for example, to a certain content being the case or to something being good; hence there is always a possibility for a practical question of whether you should be in such a state. So when it strikes me that I left my tub running and I immediately, abruptly, turn and start running back home, I can easily answer my friends who ask me what's up, without making this a case of deliberation. These cases show a sensitivity to reasons that is not the case when, for example, I have a certain sensation: I can still have it even when I know it is not the case and I have no reason to hold the state.[23] This means that

[22] See Peacocke, 1998.
[23] See Peacocke, 1998; McDowell, 1996.

I cannot unintentionally deliberate[24] but also that deliberation is a conscious activity that occupies the agent's attention.[25]

The claim that deliberation, as a process that comprises certain units or steps, is conscious means that many, if not all, these steps are present to the agent's attention, in which case deliberation would be an explicit thought process. In what follows I will explore the consequences of Korsgaard's move of equating reflection and deliberation and how this generates problems for her account of action.

1.4.1. Reflection as Deliberation

Now, there is room to interpret Korsgaard as conceiving of deliberation as an explicit and discursive process: we could go as far as conceiving of it as some form of inner dialogue that the agent rehearses prior to making a decision. In general we can think of it as an explicit thought process where all the steps are consciously rehearsed. Korsgaard's view, as noted in the quotes above, conceives of agents 'testing' their motives and this notion refers to an activity, a mental procedure through which we are directing our practical thought. It is plausible to argue that Korsgaard does conceive of deliberation in these terms because she employs the idea of the inner dialogue several times. She envisages this dialogue either between a thinking and an acting self[26] or between "parts of the soul".[27] These metaphors are explanatory of the process that Korsgaard thinks is going on during action and can support the reading that she equates reflection and deliberation.

[24] See Shoemaker, 1996, p.28
[25] Owens. 2011, p. 263.
[26] Korsgaard,1996, p.104.
[27] Korsgaard, 2009, pp.146-147. One could take these illustrations of deliberation as implying some homuncular conception of the mind but I do not think that this is Korsgaard's view. When she talks of these 'selves' or 'parts of the soul' she conceives them as functional roles within practical reasoning.

This reading raises some obvious problems: practical reasoning, understood in these terms, is a procedure that precedes action and takes time.[28] The restriction of full agency to cases of deliberation entails that those actions that are not preceded by explicit deliberation are not really "my own". Now, most of us would agree that deliberation is not so frequent a phenomenon, and most of the things we do are not preceded by this explicit thought process. This is the case for many of our daily routines and habits, as well as for many fast actions, like responding to someone's greeting, engaging in fast conversation or tackling our opponent during a football match. Nevertheless, we do not think that all those actions are not really "our own". If every trivial action required some prior explicit deliberation, from cooking dinner to having a conversation with a friend, it would be "quite hard to get through the day".[29]

Moreover, this view seems vulnerable to a regress objection. If we assume that the fact that something is done for reasons is a fundamental and necessary characteristic of actions, and once we have claimed that deliberation is necessary to set out one's reasons, we need to face the fact that deliberation is an intentional activity in its own right, and as such, it requires to be rationalized by some further deliberative act…and so on.[30] The regress also has implausible consequences for the agent's psychology, since, although "infinitely many implicit beliefs can, perhaps, be stored in a finite agent… infinitely many distinct mental actions cannot be performed at the same time by finite agents"[31]. Explicit deliberation is still a mental procedure, one that comprises steps and hence one that assumes thoughts to be discrete units that serve as such steps. This means that each step takes time to entertain and only so many of them can enter an agent's practical reasoning at one time.

[28] One issue that I do not address here is that this conception of reflection and practical reasoning is particularly prone to be threatened by Libet-like research indicating that actions are initiated in the brain before the agent's conscious intention to act.
[29] Arpaly, 2003, p. 21.
[30] See Arpaly and Schroeder, 2012, p. 220 for a detailed discussion of this regress objection. Arpaly and Schroeder highlight how this objection can be raised against different approaches to the view that deliberation is necessary for acting for reasons. They consider Korsgaard's view as an example of "Present Deliberation", a view advocating that "as we reach deliberative conclusions, we also embrace the process by which we reach these conclusions. That is, each deliberative act contains, as it were, a second act of deliberation in which the first act is deliberated upon and found reasonable—and on this basis, the fact that the first act is performed for a reason is established".
[31] Ibid.

In addition, since we are to explain agency and identity together, it seems wrong to say that identity plays a role in our behaviour only through means of explicit deliberation. As we have seen, Korsgaard usually frames the discussion of action in terms of the relation to practical identities, which means that you perform a certain action under the description of yourself as, for example someone's friend, mother etc. If reflection always involves explicit deliberation, when we truly act we would always explicitly think of our identity (as beer-drinker, breakfast-egg-eater, or even just as someone's mother, friend etc.). The picture that results from this interpretation of deliberation would seem rather implausible.

However, Korsgaard denies that this is how our identities figure in our practical thought and this can lead us to conclude that deliberation, despite the fact that it is a conscious activity, does not to require that its steps are explicitly rehearsed. In fact, Korsgaard denies that reflective endorsement needs to happen prior to the action[32] or that the deliberative process is explicit: the reference to certain reasons can be implicit, as implicit as our endorsing some self-conception.

> In claiming that such identities are the sources of our reasons, I am not claiming that thoughts about your identity need to come into your reasoning in any explicit way. But […] I want to grant that for beings like us, temptation to resist the claim of our practical identities is possible. And then you might have thoughts that explicitly invoke your identities: 'I can't do that, she is my best friend'.[33]

This suggests that the process of reasoning required for deliberation can be implicit and contemporaneous with the action. Moreover, it does not require that you rehearse thoughts about your identity or the principles guiding you. However, it is possible for you to spell these out: you can provide rather long explanations of all the features that were relevant for the action at hand. Interpreting reflection as implicit deliberation makes the model more plausible,

[32] "It is up to us to decide what justifies what, what counts as a reason for what, what is worth doing for the sake of what. We don't need to think of this, and in fact we don't think of it, as a decision made prior to action: as often as not, it is a decision embodied in the action" (2009, p. 127).
[33] Ibid., p.21.

because it allows that the steps of our reasoning are not rehearsed. It also makes it possible for the model to account for a much larger range of behaviour as an instance of full agency.

Nevertheless, I believe that an attempt to appeal to implicit, rather than explicit, deliberation is not more successful in providing a satisfactory picture of agency. While implicit deliberation, unlike explicit deliberation, need not take time since the agent is not entertaining the thought of each of steps, it still comprises some activity on the part of the agent and the regress objection still applies. Moreover, it is not clear to me what it is like for the agent to be implicitly deliberating. Above, I distinguished between acting for a reason, which still allows for our reason-giving practices, and actions that are deliberated. It is not clear to me how the phenomenology of the two differ under this reading of Korsgaard's notion of reflection: implicit deliberation seems to require some implicit rationalizing process that is not experienced and therefore does not have a distinctive phenomenal character. If so, how do we distinguish between the case in which I implicitly deliberate and resolve to buy a sweater for my mum and my mere acquiring the intention to buy my mum a sweater as in the case when this idea just abruptly dawns on me passing by a shop window. If we were to assume that implicit deliberation is actually at work in all these cases too, the difference between deliberation and other forms of thinking would be blurred. As a result we would not need to ask the third question at all.[34]

It seems that we are confronted with two rather unattractive possibilities. On the one hand, we can understand reflection as explicit deliberation: this would make reflection a rare phenomenon that implausibly excludes that we can consider many of the things we do as actions, in their fullest sense. Explicit deliberation is also affected by the regress objection: each explicit act of deliberation would need some previous explicit deliberation and so on. On the other hand, the alternative is a view of implicit deliberation whose phenomenology is unclear and that is still affected by a vicious regress.

[34] These kind of considerations are lead some to totally dismiss the philosophical interest in distinguishing kinds of agency, See Arpaly and Schroeder, 2012, pp. 228-230.

In addition, the requirement for deliberation in any form still seems to exclude all those actions that just do not involve instances of deliberation at all, either implicit or explicit. This is the case of fast and skilled behaviour, as performed in sports, or common daily actions like passing the salt to someone at our table who asked for some.[35]

Both of these specifications of reflection make Korsgaard's answer to the third question unsatisfactory. If reflection is conceived in terms of deliberation, either explicitly or implicitly, the model of agency that results from this view of reflective endorsement unduly restricts the range of action. Because the standard for full agency is cognitively too demanding, the model offers a picture of agency that is too limiting and this limitation makes the division between full actions and mere activities counter-intuitive.

The position of equating reflection and deliberation is, thus, untenable. In what follows I will discuss an alternative conception of reflection that might solve the problems I presented so far. As we shall see, this alternative does not actually help Korsgaard because, I will argue, this solution is not in fact available to her.

1.4.2. A Kantian Solution

We have seen that the implication that reflection is a kind of mental procedure inherently belongs to the language of *deliberating* and *testing* one's reasons. In response, Korsgaard can attempt to disavow this implication and deny that the agent is literally going through a process of testing. In what follows I will discuss whether Korsgaard can legitimately disavow the consequences of the vocabulary she employs to characterize reflection.

One possible solution would be to abandon the idea that her use of the term 'deliberation' means to describe the phenomenon I have defined thus far, and

[35] Similarly, not all cases of formation of belief are instances of deliberation. Opening the fridge and seeing an empty butter box, I just "see" that we are out of butter. See ibid., p. 228.

claim that her use of the term is in fact just a synonym of 'practical reasoning'. This solution would not only allow her to avoid the problems with the notion of deliberation as the process I have been describing so far, it would also lead her to dismiss the need of answering the third question about the nature of reflection altogether. If her constant appeal to deliberation entails nothing about a specific psychological process associated with a specific phenomenology, but is rather just referring loosely to practical reasoning in general, her claims are just exploring the conceptual conditions of possibility for agency. This also entails that when asserting that, in deliberating, "it is as if there is something over and above of your incentives, something which is *you*, and which chooses which incentives to act on",[36] we need to understand this claim in terms of the conditions of possibility of agency, rather than seeing it as having any bearing on the phenomenology of full actions: it is a claim about the volitional structure required by actions and about how this volitional structure is organized around rational principles.

I consider this to be a distinctively Kantian solution, and so I will now proceed to show how this proposed Kantian solution would work and why I think that it is not available to Korsgaard.

Kant serves as inspiration for Korsgaard in identifying the specificity of human agency. He thinks that, unlike animals, human beings are not determined or necessitated by their desires. This is the capacity of the human will to work on self-imposed principles which Kant calls 'spontaneity[37]. Spontaneity entails a negative definition of freedom by which the human will can operate independently from alien causes.[38] As we have seen, Korsgaard also considers that the specificity of human agency resides in the agent's self-governance and control over his desires.

[36] Cfr. Korsgaard, 2009, pp. 72, 126.
[37] Kant, Rel 6: 24; 19.
[38] Allison, 2004, pp. 136-137.

Moreover, in Kant, as in Korsgaard, we see that we identify reasons to act by incorporating the desire into one's own maxim or principle.[39] In Kant, this 'Incorporation Thesis' is meant to be a conceptual claim that plays a regulative function with respect to our conception of ourselves as rational agents: it is not an empirical claim, because, he urges, we cannot 'catch ourselves in the act of incorporating desires into our maxims'[40]. This means that incorporating desires in maxims is not an object of possible experience that we can introspectively access. It also is not a metaphysical claim, since Kant denies that it is meant to bear on the reality, or even the real possibility, of the kind of freedom it assumes.[41] The claim is thus compatible with the possibility of us not being more than complex mechanisms rather than spontaneous rational agents,[42] so that I might be deluded in believing that "my reason has causality".[43] Moreover, we do not get to understand our capacity of spontaneity through the experience of the incorporation that is necessary for our agency: in Kant it is the experience of our moral duty that allows us to understand the freedom from determination of all other desires and informs us of the independence of our will.[44] So the claim is not that we are free only inasmuch that we act from duty, but rather that our capacity of acting from duty makes us aware of its conceptual requirement and hence of the spontaneity of our agency. As a result, within the Kantian picture there is no mention of the mental processes practical reflection employs, and phenomenological concerns have no room in Kant's attempt at determining the conceptual conditions of practical reason.

Following Kant, Korsgaard could dismiss the objections raised above by arguing that 'reflection' is not to be understood as describing any experientially accessible process. However, I don't believe that Korsgaard has this Kantian solution open to her, for two reasons: firstly, the Kantian solution cannot serve

[39] Kant, Rel 6: 24; 19.
[40] Allison, 2004, p. 118, 134; Kant, Reflexion, 5661, 18, pp. 318 – 319.
[41] Kant, *Critique of Pure Reason*, A 557 – 58/B585-586.
[42] Allison, 2004, p. 124.
[43] Ibid., p. 134.
[44] Ibid.,p. 139. Here I take on board Allison's interpretation of the moral law as the *ratio cognoscendi* of freedom, which is based on the following passages of the Metaphysics of Morals and the Critique of Practical Reason where Kant says that the moral laws "first make known to us a property of out choice [Willkür], namely, its freedom" (MS, 6:225:51) and that apart from our consciousness of standing under the moral law "we would never have been justified in assuming anything like freedom (Kpr V 5: 4n; 4).

Korsgaard's constitutivist purposes; secondly, there is evidence that Korsgaard's characterization of reflective endorsement does have experiential commitments by grounding her model in the agent's *identification* with his principles. To show this, I will first highlight the different aims that Korsgaard's and Kant's projects seem to have.

Unlike Korsgaard, Kant does not attribute agency exclusively to autonomous agents and does not aim at distinguishing mere activities and full-blown actions. The difference between the two philosophers is immediately clear if we consider their very different treatment of heteronomous agents (who, in Korsgaard's terminology, can be equated with wantons). In Kant, heteronomous agency is not some sort of lack of agency or a lesser degree of agency,[45] as it seems to be in Korsgaard account. Views such as Korsgaard,

[45] One possible response that I do not address directly here is the fact that, given that according to Korsgaard the categorical imperative is the constitutive principle of action, heteronomous principles are defective as principles, hence the kind of behaviour they guide is defective, both morally and in terms of their degree of agency. Lesser degrees lack agency since they are based on some defective principle. I believe this suggestion is problematic but the detailed examination of these problems would lead me to discuss Korsgaard's interpretation of Kant to an extent that would not be compatible with the aims of this work. Because of this I do not discuss this point further in my main discussion, but I will sketch my concerns here. Remember that Korsgaard (2008, p. ix) describes the idea of being guided by a principle in the following way: in order to engage in a certain activity (building a house, swimming, running etc.) you need to be guided by a conception of what the activity amounts to (for example, moving your body in ways so that you can proceed forward in the water). Such a conception is the constitutive principle of your activity and sets the standards for that activity to be what it is. This means that constitutive principles are at once descriptive (inasmuch as an object must meet them to be the thing/activity) and normative (since an object can fail to meet these inner conditions and thus been considered defective). On the other hand, the application of constitutive principles comes in degrees: it is not true that you are not performing an activity unless you are performing it precisely. It is true, though, that to perform an activity at all you must be guided by some precise version of it. So if you are trying to swim, you are still guided by the constitutive principle of swimming inasmuch as that is the activity you want to perform (and you are not just splashing around). This does not mean that your performance cannot fail to correspond to the perfect application of the principle, and indeed you might, in the performance, just be splashing around. Only because you are failing in performing the activity though, I am allowed to say: "You are swimming very poorly". Korsgaard employs a Platonic notion to explain how principles work: your performance *participates* at different degrees with the perfect form of the activity. This means that, in acting, we can establish degrees of participation to the categorical imperative. I find this idea problematic, and I am unsure whether or not it is possible to apply the idea of degrees of participation to a principle like the categorical imperative. In cases of defective performances in activities like swimming, walking etc., it is true that the principle is still at work and hence makes it possible to have a normative judgment on the performance. Someone who is trying to swim and ends up just splashing around is still applying the same principle as Michael Phelps does; he just cannot perform very well. In the case of action though, we cannot say that the intermediate forms of unity and agency are still applying the categorical imperative, just not very well. Flawed forms of unity are not failing to apply the categorical imperative; they are guided by some other principle altogether. I will not explore further the possibility for there to

which do hold this kind of understanding of heteronomous agency, do not make sense of Kant's criticism of moral theories that are based on principles other than the moral law, such as those that are based on self-love. He thinks that the subjects endorsing them are heteronomous, since their principles fall short of the requirements of rationality embodied in the moral law. Nevertheless, Kant presupposes that these heteronomous subjects are agents: they act on the basis of principles rather than merely respond to stimuli.[46]

Korsgaard acknowledges[47] this fundamental difference between her view and Kant's on this point and her solution invokes the vindication of the categorical imperative as a constitutive principle of agency. Whether or not her solution is successful is not relevant here. What matters here is rather that Kant's picture does not attempt to make claims about the metaphysics of agency, and it does not need to appeal to Constitutivism in order to provide us with a taxonomy of action. Because Kant does not share the same constitutivist aims as Korsgaard's, the fact that the Incorporation Thesis does not provide a psychological and phenomenological account of reflection is not a problem for his model. In this light, though, we must notice that Kant's response to the sceptic about morality is distinctively weak, since transcendental freedom is entirely unknown and unknowable, and since we have seen the idea that we are free could potentially be entirely illusory.

Constitutivism finds its strength in the fact that the descriptive standards of action allow us to identify the normative standards of good actions. This assumes that there is something these standards are descriptive of. In the Kantian picture,[48] the reply Constitutivism can offer to the sceptic is a rather odd one, because there is always the possibility that there is no such a thing as true 'action', as Korsgaard models defines it. There might be nothing that

---

be degrees in the exercise of agency in respect of the categorical imperative, because the argument in the present paper does not need this further discussion and because analysing the notion of participation would require me to expand further on aspects of Korsgaard's theory that I have not presented fully.

[46] Allison, 2004, p. 135.

[47] Korsgaard, 2009, p. 81.

[48] This is true regardless of how we interpret Kant's transcendental idealism, whether we are giving a reading of it in terms of double-aspect or double-object theory. For a distinction between these two interpretations see Allison, 1983.

corresponds to these conditions identified conceptually, and no metaphysical category to which those conditions apply. This is why I believe that this Kantian solution does not fit with Korsgaard's purposes.

The second point that questions the possibility that the Kantian solution can be of much aid for Korsgaard's project is evidence that Korsgaard's characterization of reflective endorsement essentially includes elements about the agent's experience that are alien to Kant's picture. Because of this, I think that when Korsgaard claims that when you are acting, "it is *as if* there were something over and above your incentives', she makes explicit reference to the experience of agency.[49] As we have seen, in the Kantian solution, this would not have been possible.

If this is the case, Korsgaard, unlike Kant, needs to provide an account of reflection that can explain these elements for which a generic account of reflection as practical reasoning cannot suffice. One such element is the agent's *identification* with the principle that is required to attain the volitional unity. I believe that the importance of the notion of identification for Korsgaard can be made clearer if we see how the model handles certain problematic cases, highlighted by Arpaly:

> Huckleberry Finn. "Huckleberry's best judgment tells him that he should not help Jim escape slavery but rather turn him in at the first available opportunity. Yet when a golden opportunity comes to turn Jim in, Huckleberry discovers that he just cannot do it and fails to do what he takes to be his duty, deciding as a result that, what with morality being so hard, he will just remain a bad boy (he does not, therefore, reform his views: at the time of his narrative, he still believes that the moral thing to do would have been to turn Jim in).[50]

In this case of inverse akrasia[51] Huck is failing to acknowledge that his acting is responsive to the fact that Jim is a person, a human being just like him (in Kantian tones we could say that he is responsive to the principle of humanity).

[49] In Korsgaard, 2009, p. 126, we can see this claim directly associated to the phenomenology of agency.
[50] Arpaly, 2003, p.9.
[51] The case is described for the first time in Bennett, 1974, and then extensively in Arpaly and Schroeder, 1999.

He believes that he endorses slavery and is baffled by his own acting. He finds that he is psychologically unable to do what he believes is the right thing and conceives of himself as weak-willed. Moreover, it is not just that the guiding principle is obscure to Huck, the very maxim that describes his action is: Huck fails to acknowledge that he is 'helping a friend' and the description 'helping a slave' is not something he can endorse, in the light of what he takes his convictions to be. In this sense he seems to be unaware of his reasons.

Moreover, rather than deliberating about his action, Huck seems to be acting from instinct and therefore it cannot satisfy the conditions for agency in Korsgaard's model. Here is Korsgaard's characterization of the possibility of acting from instinct:

> The experience of acting from instinct is obviously not, phenomenologically, like the experience of applying a rational principle to a case. But for that matter, acting on a rational principle need not involve any step-by-step process of reasoning, for when a principle is deeply internalized we may simply recognize the case as one falling under the principle, where that is a single experience. Principle and instincts play a role in structuring our perceived environment.[52]

Here Korsgaard claims that one is not to first adopt some principle, and then select some more specific maxim so that this requires an explicit choice, but it is required for the principle to be recognized and presupposed by one's more specific maxims as their ultimately explanatory ground.[53] Notice that these remarks do not offer any precise indication of what the reflection necessary for this recognition of one's principles amounts to; however, not even this very vague form of accessibility and recognition can be ascribed to Huck.

Because Huck fails to acknowledge that he is in fact acting in light of the value of Jim's humanity, it seems as if this is a case of 'unreflective endorsement', where a principle[54] is operating and guiding the action despite the fact that it does not provide the explanatory grounds for it. If we do not assign any

[52] Korgaaard, 2009, p.107.
[53] This is a view that can also be found in Kant, cf. Allison, 2004, p. 116.
[54] I am assuming Korsgaard's view of acting as necessarily relying on principles. I am not exploring here how the kind of cases presented could serve to show that we should dismiss this whole idea. See Arpaly, 2003, for arguments in this direction.

essential role to the access and the experience an agent has of his principle, it could seem that in 'structuring the received environment' an internalized principle could be something that I do not acknowledge or I could even deny having. If the experience was completely irrelevant and the process of reasoning and being guided by a principle completely below the threshold of consciousness, we could consider as instances of agency cases of agents living truly unreflective lives, who even deny experiencing agency but whose actions follow the categorical imperative. However, if this were the case, it is unclear what the *reflective* in "reflective endorsement" stands for. These cases point at the fact that the model needs to specify the conditions for *reflective* endorsement, rather than other possible forms of commitment.

We have seen that in Korsgaard's model the function of the principle constitutive of agency is to produce the volitional unity necessary for the agent to act at all. If this role could be played by an unreflected and even unacknowledged principle, we could ascribe volitional unity to Huck Finn, and see him merely lacking the access to his guiding and unifying principle. However, this is not how Korsgaard conceives the unity necessary for agency: that unity requires identification. In her view, identification with a principle plays a role in structuring the agent's will which is not devoid of an experiential component.

> When you determine yourself to be the cause of the movements which constitute your action, you must identify yourself with the principle of choice on which you act. For instance, suppose you experience a conflict of desire: you have a desire to do both A and B, and they are incompatible. You have some principle that favours A over B, so you exercize this principle, and you choose to do A. In this kind of case, you do not regard yourself as a mere passive spectator to the battle between A and B. You regard the choice as yours, as the product of your own activity, because you regard the principle of choice as expressive, or representative, of yourself.[55]

Identifying with some motives or feeling alienated from others seems to be a feature of the agent's experience: one that contrasts the alleged passivity of a mere bystander to the activity which is proper to agency. When an agent does

[55] Korsgaard, 2009, p. 75.

not experience that he endorses his motives, he is passive in respect of desires that determine him. In this sense, Huck is alienated from his principle: he does not identify with it. For Korsgaard, identification is constitutive of agency because, as we have seen, one's identity is an activity: it is an activity that requires that one does not feel estranged and alienated in respect of one's desires. Moreover, identification is fundamentally linked to the role played by practical identities. The reference to practical identities can make it even clearer that experience is central in Korsgaard's picture:

> An agent might think of herself as a Citizen of the Kingdom of Ends. Or she might think of herself as someone's friend or lover, or a member of a family or an ethnic group or a nation. She might think of herself as the steward of her own interests, and then she will be an egoist. Or she might think of herself as the slave of passions and then she will be a wanton. And *how she thinks of herself* will determine whether it is the law of the Kingdom of Ends, or the law of some smaller group, or the law of egoism, or the law of the wanton that will be the law that she is to herself.[56]

As we have seen, thoughts about one's identity need not figure in any explicit way.[57] Accordingly, I can think 'I need to feed my cat' or 'I am going to buy the best Parmesan', even if the explicit self-attributions 'I am a cat owner' or 'I am a foodie' seldom cross my mind. Nevertheless, as we have noted above already, 'it is not that you have a personal identity you might or might not be conscious of. If you are not conscious of it, than you don't have it'.[58] Even with these limitations on the explicit content of the agent's thought, this kind of endorsement requires an object, one that is related to the content of the agent's thoughts. The point is not that such a content must have some explicit reference, but that the agent's activity, as reflection upon this content, makes room for demanding an account of what such reflection amounts to. I will develop these considerations further in chapter 3, what is important to stress here is that the appeal to self-constitution as a self-conscious activity the agent engages in calls for an explanation of such an activity and the Kantian picture, as we have seen, denies that it is at all possible to do so.

[56] Korsgaard 1996, p.100 (my italics).
[57] Korsgaard 2009, p.21.
[58] Korsgaard 2008, p.214n.

It may be objected that the way I characterized identification might seem ambiguous: on the one hand I stressed how identification is a feature of the agent's experience; on the other hand I described it as a psychological process that determines the structure of the agent's will. Conceptually, these are quite distinct things.[59] I believe Korsgaard's notion of identification does contain this ambiguity and I will discuss in chapters 3 and 4 how her conception of agency rests on problematic assumptions about the nature of the self and self-consciousness. In those chapters, I will highlight how these assumptions eventually generate the problems I have discussed here. All that it is important to notice at the moment is that Korsgaard, unlike Kant, does associate an essentially experiential component to her notion of reflection. So, regardless of how Korsgaard can reconcile the two conceptually distinct aspects of her notion of identification, there is compelling evidence in support of the claim that Korsgaard does include features of the agent's experience in defining her conception of reflective endorsement. If this is the case, the Kantian solution that sees 'reflection' as merely a synonym of practical reasoning leaves vague and undetermined the conditions of endorsement. Either phenomenology matters and needs to be explained by giving an account of the reflection implied by reflecting endorsement, or it does not matter at all, and then there cannot be any appeal to it and it has no bearing on judging whether or not some behaviour counts as action or not.

1.5. Conclusions

To sum up, Korsgaard characterizes reflection as deliberation. If we take this literally we end up with a counter-intuitive and restrictive picture of what counts as action. If we take deliberation to be an explicit thought process, Korsgaard's conception of agency is just implausible. If we take it to be an implicit process it still seems to be too restrictive and it also becomes a phenomenon with a very unclear phenomenology. If we take it to be just a

[59] See Arpaly and Schroeder for a similar discussion on the role played by 'alienation' in Frankfurt. As they remark, "an experience of a desire as having certain features is something quite distinct from a desire actually having those features" (1999, p. 379).

synonym of practical reasoning we face two difficulties: firstly that it is insufficient for the aims of Korsgaard's Constitutivism; secondly, that the theory provides only a vague account of what reflection amounts to, with the result that it is questionable if it can have any explanatory power at all in establishing whether some behaviour counts as action or not. In fact, because of the importance of the notion of identification, Korsgaard cannot dismiss, in Kantian fashion, the relevance of phenomenology. The notion makes the agent's experience an essential element of her model in a way that it is not in Kant's. I believe that there is a reason for the fact that the theory gives us an unsatisfactory account of reflection, and I will trace this in the assumptions that underlie her notion of the self as well as the notion of control and guidance. I will address these concerns directly in chapter 3 and 4. Before that, I shall examine David Velleman's model. As we shall see, Velleman's model, although committed to the view that reflective endorsement plays a constitutive role in agency, has some features that appear to avoid the difficulties that emerged in this discussion of Korsgaard's model. However, I will argue that his model is not satisfactory and it is, in fact, subject to some decisive objections.

Chapter 2

Velleman's Narrative Model of Agency

In the previous chapter we have seen that the requirements for full agency presented by Korsgaard's model are unsatisfactory. In particular, the prominence assigned to deliberation and the demands imposed by the categorical imperative as a test for endorsement lead to an unduly restricted range of what we can consider actions. This criticism leaves open the possibility that we can preserve the aims of Constitutivism by employing a different account of the test for full agency free from these charges. David Velleman's narrative model is a promising candidate to achieve this: he is explicitly wary of making deliberation necessary for practical reasoning and his appeal to narratives in defining the conditions for endorsement aims at providing a model that is not cognitively as demanding as Korsgaard's. In line with other theories that focus on the role of reflective endorsement in agency, Velleman claims that there is a relevant distinction between two kinds of activities human agents can engage in: motivated activities, which some non-human animals can also perform, and full-blown actions, which are specifically human. For Velleman, behaviour that counts as action springs from the agent's self-governance and from his control of his motives. This means that Velleman's model involves two kinds of agency and the central feature of actions (in contrast with motivated activities) is the agent's *authority* over them.

Velleman's model conceives reflective endorsement in terms of the action fitting with one's own narrative understanding. I will label Velleman's a *narrative model of action*, since he characterizes acting as *enacting* a coherent

narrative. Actions are, in this view, partially motivated by our understanding of them under certain descriptions and this understanding places actions in the framework of a coherent narrative that qualifies one's motives as reasons to do them.

Velleman's view that narrativity is constitutive of human agency is grounded in the intuitive idea that our capacity for engaging in narratives sets us apart from other creatures. Some[1] have gone as far as to say that we should be named *Homo Narrans,* rather than *Homo Sapiens*. From the idea that we acquire a basic understanding of folk psychology for the first time in ontogeny by engaging with narratives, to theories that identify human selves with narratives, narrativity has featured in a wide variety of philosophical debates.[2]

There is a natural connection between story telling and reason-giving since it is often the case that in everyday life we make sense of what we do in the form of (sometimes very brief, sometimes very articulated) stories, which involve our reasons to do what we do. It then seems that narratives can play a central role in understanding human agency.

In the first part of this chapter I will present how the narrative model of agency answers the three questions I established as central for any theory of reflective endorsement. I will consider how these answers are connected in Velleman and then I will highlight four features that ground the possibility for narrativity to play the role the model assigns to it. I present Velleman's narrative model of action covering material that is spread throughout different collections of articles. I consider that, despite some revisions, Velleman's work on narrativity is of a piece with his previous work on practical reasoning. Part of this chapter is therefore devoted first to presenting the narrative model and then to evaluating the role assigned to narrativity in his wider picture of

[1] Neimeyer, 2003.

[2] In fact there is an extensive literature on the topic. As we shall see in chapter 3, this includes substantial work done in cognitive science and psychology. In philosophy, the range of authors interested in the role of narratives crosses the boundaries of continental and analytic philosophy, being an interest of people working in phenomenology as well as action theory of theory of mind. Cf. Hutto (2007), Dennett (1991), Gallagher (2003), Ricoeur (1988).

agency.[3] In the second part of the chapter I will present some of the attractions of the narrative model as well as some advantages for Velleman's narrative account over other theories that also stress the importance of agential authority, such as Korsgaard's. In the last part of the chapter, I consider two objections. First, I consider how cases of self-deceiving agents constitute a substantial problem for the narrative model: Velleman's criterion for agency is too demanding and unduly restricts the range of actions that the narrative model can account for. This problem originates from the implausibly high standards for self-understanding that the theory relies on, and I will show how these are entailed in Velleman's conception of narrativity. On the other hand, I will question the compatibility of Velleman's two characterizations of the reflection required in action with the role assigned to the self as a narrator.

2.1. The narrative model of agency

Velleman's narrative model of agency is built on three theses. I will give a full account of it by examining how these thesis answer our three original questions. I shall start with the first:

2.1.1 *What is the relation between self-consciousness and self-governance?*

(i) The 'two dimensions of agency' thesis: The central intuition of reflective endorsement theories of agency is that the human capacity for self-consciousness gives humans a special relation to their behaviour: they can step back and call into question their actions, as well as their motives and attitudes. Furthermore, the agent *commits* to some of these motives, regarding them as reasons and thus performing the relevant behaviour in light of them. Thus, because of our cognitive reflective capacity, our actions are not mere outcomes of whatever state we are in. Velleman contrasts actions with slips or unconsciously motivated movements: my breaking the old inkstand because of

[3] I take Velleman's to be a unified and coherent project that therefore allows for this attempt. Velleman himself recognized gaps still in need of filling in his reply to Catriona McKenzie; see 2007b, p. 284.

a desire for my sister to buy me a new one is a motivated doing. However, my behaviour is not guided by my motive *as a reason* and indeed in the performance I am not even aware of this desire of mine. Thus, for Velleman, this doing counts as a slip rather than an action. On Velleman's conception of reflective endorsement, motives can count as reasons only when the subject reflects on them and *reinforces them by self-understanding*.

It is important to notice that Velleman's model has two starting points: one is the endorsement of Davidson's project of identifying that "causal mechanism that has the function of basing one's behaviour on reasons";[4] the other one is the view that Davidson's picture is insufficient: standard causal theories fail to distinguish the two dimensions of agency, since motivated activity of the sort described by Davidson can be attributed to some lower animals and does not capture the specifics of human agency.[5] In order to improve on Davidson's model, Velleman says, there must be something we need to add to 'bring the agent back into the picture' and, if we are to hold on to the reductive inspiration of the causal theory, there must be some psychological element that functions as the agent and plays its role.[6] It is in order to fulfil this role that Velleman introduces an 'intelligibility drive':[7] this drive does not merely push one to know what one is doing but also why one is doing it. Behaviour that 'makes sense', in these terms, is reinforced by the motivation associated with this drive. Lacking an understanding of what I am doing, I stop at the kitchen door and I wonder 'What did I come here for?'. Velleman urges that in instances like these, we are explicitly aware of this intelligibility drive and we are inhibited from executing actions that we do not understand. One's self-knowledge reinforces certain courses of actions rather than others because people have a motive to do what makes sense to them. When I judge that 'yes, all things considered, ϕ-ing makes sense' my motives of self-understanding strengthen the desire that motivated me in the first place and as such can be

[4] Velleman, 2000, p. 6.
[5] Velleman, 2000, pp.10-11.
[6] Ibid., p. 6.
[7] By which he means a tendency to make sense of what we do and what goes on around us but which, unlike our reasons, does not need to figure among our conscious motives. This distinction is explicitly stated in the new introduction to *Practical Reflection*, in which Velleman still speaks of a 'desire', rather than a 'motive', for self-understanding. See Velleman (2007a) p.xx

causally efficacious, in much the same way standard causal theory assumes. Practical reasoning therefore surveys the agent's motives, inhibiting some of them and reinforcing others. Ultimately then, reasons can be causes because the intelligibility drive reinforces the original motivation to perform those actions associated with motives that count as reasons.

2.1.2 *What is the "self" of self-governance?*

(ii) The 'two dimensions of selfhood' thesis. Velleman thinks that a conscious animal has a first-person perspective but no representation of itself as occupying such a perspective. We humans have this but, in addition, we also have an objective conception of the creatures we are. The capacity for reflexive self-awareness is the capacity for shifting from a first-personal to a third-personal perspective: the subject makes oneself the object of reflection. Velleman conceives of this objective self-representation as a story: in exercising his capacity for self- consciousness, the agent places his motives and his actions within a narrative framework that is necessarily associated with a certain self-description. This picture gives us two dimensions of selfhood: a minimal, purely first-personal perspective, and a narrative self-conception.

2.1.3 *What does reflection amount to?*

(iii) The two kinds of reflection thesis. Velleman emphasises a distinction between the actual occurring thoughts the agent engages in and the logic of practical reasoning.[8] An intention to act need not be a reflective conclusion that emerges in articulated thought, nor does there have to be some sort of reflective monologue taking place in the agent's mind. Reflective intelligibility relies on background self-ascriptions that implicitly convey the attitudes that an explicit thought would express. Though unarticulated, these self-attributions serve as reasons for acting because they provide the logical structure of the agent's thinking.[9] In this sense, we can posit a kind of reflection that must be distinguished from deliberation and would mainly be *supervising* the stream of

[8] See Velleman (2007a), pp.88-89 and 106-107.
[9] See Velleman (2006), p.281.

our behaviour already underway. I will call this kind of reflection 'perceptual' because it does not involve the agent as explicitly reflecting prior to the performance of his behaviour, but rather it just involves the agent's perception of the circumstances. This kind of reflection is something in between automaticity and deliberative reflection. In getting up and almost mindlessly preparing yourself a coffee you perceive both your movements and the present circumstances, adjusting the stream of your behaviour to them. In the process, practical reasoning is supervising your conduct, placing your behaviour under action-concepts that make sense of it. As a metaphor, we can understand practical reasoning as occupying the passenger's seat, rather than being the actual driver, or as a supervisor, rather than the actual labourer. The supervisor, through his understanding, follows along the stream of behaviour, making sure that it fits, that things are on track; and it does so by mostly just staying in the background (both of action and thought).[10]

To sum up, according to the picture of agency provided by thesis (i), full-blown actions are those that are guided by this intelligibility drive and are hence regulated by reflective understanding. Once the mechanisms leading to action are in place we still need to establish the condition for something to "make sense". This is where narratives play a role. Thesis (ii) defines the terms of endorsement: the condition for something to be endorsed is that it passes what Velleman calls a *coherence* test, against the background of one's own narrative self-conception.[11] In testing his motives against this background, the agent can identify his competing motives and throw his weight behind some, reinforcing them via his endorsement. Minimal self-reference is insufficient for full actions, given that my self-conception needs to function as a coherence test by which my acts, under certain descriptions, can be judged as fitting with a certain narrative or not. A narrative involves self-description (and self-ascription) and as such it needs to cohere with both our other stories (past and future narratives) and the constantly revised evidence that contributes to our

[10] One might find this picture confusing since it seems "shaping the driver's understanding" is just not the same function as "keeping the driver on the right track". As we shall see this legitimate concern is answered once we establish the connections that Velleman draws between control and understanding.
[11] Velleman, 2006, p.218.

understanding of the world. Coherence constrains the range of actions available to us. According to this picture, the higher (full) level of agency is possible only when a subject acts through such self-representation: narrativity is characteristic of all full agency. Thesis (iii) completes the narrative model by characterizing the kind of reflection at work in practical reasoning, while also describing the narrator not just as a deliberator, but also as a supervisor.

It is important to stress how the narrative model of action is conceived by Velleman to be an improvement of Davidson's causal theory of action, rather than an outright rejection of it. As we saw, Velleman believes that Davidson's model is insufficient:[12] it does not adequately distinguish between actions and mere activities and it does not explain how our motives can figure as reasons in causing behaviour. However, he wants to find the elements within our psychological makeup that play the role of the agent, while retaining the same causal structure.[13] Such elements are the "intelligibility drive", which alters the overall balance of one's motives and thus directs one's actions, and a "self-consistency drive", which preserves the coherence between behaviour and self-knowledge. Because of the latter, we have the tendency to act in accordance with our self-representation and we make our behaviour cohere with our narrative. When a person thinks of performing a certain act that fits with her self-interpretation, she does so attributing a certain act-description to her action. And when she acts, she instantiates it. Actions are the fulfilment of the agent's narrative self-conception. So for example, a child is in the position of choosing whether or not to lie to his parents and skip school to be with his friends. [14] According to Velleman, only inasmuch as he can somehow coherently fit the lie within his self-conception will he be willing to tell it. The self-governed agent is conceived as a narrator enacting his own autobiography.

[12] Velleman, 2000, p.10-11.
[13] Ibid., p. 6.
[14] 'If the boy finds a way to reconcile the lie with his self-conception – a story to tell himself about telling the lie, which would amount to a rationale for telling it – then his practical reason condones telling a lie and he is consequently "willing" to tell it' (Velleman, 2006, p. 251).

In order to complete the picture given by the three theses above, there are four features of narrativity that we need to consider.[15] These four features enable Velleman to build his narrative model of action, and his metaphor of the self as a narrator is bound to them. The first one is "authorship". In this Velleman departs from Dennett's famous formulation of the narrative self: authorship is the key to denying the purely fictional character of the self, as favoured by Dennett, and accounting for how the agent's reflective representation can be actually productive of behaviour. Dennett's main aim in describing the self as a "center of narrative gravity" was to get rid of the idea that it could be a "central meaner or controller".[16] He concluded that: "Our tales are spun, but for the most part we don't spin them; they spin us. Our human consciousness, and our narrative selfhood, is their product, not their source".[17] According to Dennett the narrative self is the fictional protagonist conjured up by the organism, which is the real author, to serve the function of providing coherence to our lives. This self gives us the illusion of being actually in control of our behaviour, when in fact it is just a result of our fictionalizing tendencies. Velleman criticizes this view and urges that Dennett's account fails to explain phenomena like Multiple Personality Disorder[18]. In the discussion of these cases Dennett conceived different modules in the brain as working and processing various kinds of information in an essentially independent way, although these different subsystems could mutually influence one another. Velleman points out that Dennett's account fails to account for the fact, when 'switching from one self' to the other, the patients' behaviour changes in a variety of aspects, for example the way they talk or walk, which seem to belong to the specific personality that each self carries along. Velleman notices how in Dennett's picture it becomes difficult to understand why a patient does not ever speak according to one self and walk according to another. Such elements of behaviour, he concludes, must be interdependent, because the self, acting as a functional mechanism, coordinates them.

[15] I am not here claiming that Velleman presents these four features when characterizing his idea of the self as a narrator. In fact, when presenting the salient characteristics of narratives in his "Narrative Explanation", Velleman mentions the following three: coherence, meaningfulness and emotional import. However, the following are characteristics that can be inferred from his work contained in Velleman, 2003, but also in Velleman, 2006.
[16] Dennett, 1991, p.422.
[17] Ibid. p. 418.
[18] Ibid. pp. 419-420.

Velleman thinks that the view rejected by Dennett – identifying the author and the protagonist – is in fact correct. Due to a self-consistency drive, Velleman's agent strives to maintain coherence with his story and can (1) literally decide which path to follow and (2) take himself to be the author of his deeds. The agent's self-conception thus feeds back into behaviour and it does so by means of actual agential control, and in this sense the narrative plays a role that is not just "fictive, but factual".[19] The shift from considering the narrative a biography of a purely fictional character to defining it an autobiography reinstates the agent in the picture with an autonomous causal role and conceives of agency in terms of authorship. The autobiography and the behaviour it narrates are mutually determining.[20]

Secondly, narrativity conveys meaning: this allows the act-descriptive coherence that belongs to narrative to provide the endorsement test. A narrative explanation is not a mere causal account of the events that led to a certain conclusion; it reveals the connections of thoughts, feelings and actions[21] in light of which the narrative makes sense. Meaningfulness is crucial when it comes to form the act-descriptions that are the subject of reflective scrutiny. My story is thus grasped under related concepts (e.g. betrayal, celebration, etc.) that convey their significance. In particular, it involves my self-descriptions, so that something makes sense to me inasmuch as I am a friend, a citizen, a woman etc.

A third element, which Velleman stresses in order to contrast his view with Dennett's, is the *episodic* character of narratives.[22] Velleman conceives of several small narratives, as small as cooking a meal, sending a text or having a coffee break. The idea that narrativity is episodic, combined with the dimension of meaning given by the agent's self-conception, makes narratives provisional, tailored to the situation and to the agent's understanding of it; moreover, it

[19] Velleman, 2006, p. 221.
[20] Velleman, 2006, p.211.
[21] For a similar example see Velleman, 2003, p. 10. Here Velleman also emphasises how narratives convey emotional understanding; see ibid,. pp. 18-19.
[22] For an objection see, for example, Mackenzie, 2007, pp. 263 – 282.

allows Velleman to deny that we need to consider a person's autobiography as a single unified story that incorporates all of the minor episodes.[23]

Finally, Velleman argues that narratives need not be actually formulated in discursive terms and need not be fully determined in all their details. This means that narratives allow some degree of *vagueness*, due to both their being meaningful and provisional: the meaningful context in which they fit enables the agent to skip a full account of all details. Narratives build a character, a scenario,[24] in which not all elements need to be told to function as an explanation and a setting. However, the fact that not all elements appear in the explicit narrative does not entail that they do not play a role implicitly in the background. This means that the narrative/supervisory self can work implicitly, as perceptual reflection requires.

2.2. The attractions of the Narrative Model

The narrative model has considerable attractions: it not only develops the intuitive idea that narrative practices are pervasive and distinctive of human agency, but it also has advantages over other theories that ascribe a constitutive role in human agency to agential authority such as Korsgaard's. I shall explain what these are.

First, narrative coherence does not seem to be too demanding a test and, in combination with thesis (iii), it allows the narrative model to account for a wide range of actions which, since they are not instances of deliberation, were necessarily excluded from Korsgaard's picture. This is particularly important, since Velleman explicitly aims at giving an account of practical reasoning that

[23] This means that the issue concerning the conditions for agency can be functionally distinguished and considered separately from concerns about personal identity and persistence of the same self through time. Moreover, notice that the idea of plenty of small narratives allows that there can be conflicting (from someone else's point of view, as well as from ours once the connection is pointed out) idiosyncrasies among these units.

[24] See Velleman, 2007b, n.6, where Velleman explicitly prefers to refer to a scenario instead of a script, given the higher degree of specificity required by the latter and the greater indeterminacy allowed by the former.

is not overly intellectualistic:[25] the appeal to narrativity might then seem promising in addressing the charges that his view is too cognitively demanding.[26] As we shall see this leads Velleman to think of practical reasoning as 'a peculiar mode of theoretical reasoning'.[27] Even when engaging in perceptual rather than deliberative reflection, practical reasoning is the author of the agent's behaviour: much like a supervisor who is responsible for the results of the workers he supervises. Velleman's distinction between the logic of practical reasoning and the explicit content of practical thought allows the reference to one's own self-conception posited here to be implicit.[28] I believe that this allows us to understand the supervisor as credited with a form of narrative intelligence: his narrative guides behaviour by means of what I shall call an *implicit narrative structure* (INS)[29] that shapes the agent's own understanding. This INS provides a test for coherence through which one endorses one's motives and doings. This means that even the most mundane and almost mindless acts, like getting a beer from the fridge, or one's bedtime

[25] Velleman, 2007b, p. 284. Velleman introduces the reference to narratives with this explicit purpose.

[26] Ibid. I take it that with thesis (iii) Velleman also aims at completing his picture of practical reasoning thus avoiding the charge of confining his analysis to "paradigmatic cases" of action. See Velleman, 2004b, p. 282. Here Velleman replied to one of Mele's objections that "The human capacity to chew gum may raise interesting questions, I suppose, but they are nothing compared with the questions raised by the capacity to deliberate between alternative futures and to bring one of them about for good reason"; it seems that with (iii) and his appeal to narrativity, his model of action tries to accommodate even these previously neglected cases.

[27] Velleman, 2007a, p. 90. According to Velleman, the agent engages in a kind of reasoning, which while being indeed theoretical, does not alienate him from his actions leaving him "as a bystander or a commentator" (ibid., p.92).

[28] This is necessary, since the very idea that our understanding provides us with an act-description that is at the same time a self-description, could lead to the objection that acting out one's own self-conception (as enacting one's own narrative) would turn agency into a rather egocentric enterprise. It does not seem to capture the way in which one's attention is focused on the outside world and on the other, rather than on oneself. It seems implausible to assume that the agent's thoughts are always somehow concerned with "who he is". When helping out a friend in distress, my thoughts will most likely be about her, about the best ways to help her etc. rather than about myself acting as a true friend etc. (Although, of course, sometimes this would be the case. For example when I am torn between different things I care about. I can dismiss my work duties or previous arrangements and think "I cannot do that, I am a friend of hers and as such I cannot abandon her!"). Velleman addresses this issue directly both in the "Appendix" of *The Centered Self* and in the "Introduction" to *Practical Reflection*. See Velleman, 2006, p. 281; Velleman, 2007a, p. xxviii.

[29] Note that talking of practical reasoning as having an implicit narrative structure does not need to introduce the idea of some prior script that the supervisor follows: "the self-narrating agent is a bit like an improvisational actor, enacting a role that he invents as he goes", (Velleman, 2006, p.221). The supervisor simultaneously with behaviour frames certain act-descriptions, making sure that we act *in character*.

routine, are actions in the fullest sense, because they are supervised to fit an INS.

Another good feature of Velleman's narrative model is that it seems to me to stand against some criticisms raised by Galen Strawson to narrative views of the self.[30] Strawson argues that "if someone says, as some do, that making coffee is a narrative that involves Narrativity, because you have to think ahead, do things in the right order, and so on, and that everyday life involves many such narratives, then I take it the claim is trivial".[31] Since Velleman's narrative model aims at accounting for such instances of behaviour, like my morning routine of coffee-making, I want to briefly say something about how it is possible for the model to withstand the charges of triviality. After all, the threat of triviality might seem to be particularly worrying for Velleman: as we have seen, his interest in narrativity is linked to short-term, present contexts and conceives of narratives as episodic and provisional. Therefore he cannot appeal to narrativity in order to embed such episodes in extended, potentially life-long, unified narratives, which is something that would be necessary if, for example, one was to hold substantial claims about one's life being structured by the unity of one's individual narrative.[32] Moreover, the very idea of INS presented above could seem to allow us to take narrativity as nothing more than some form of understanding, explanation or description. Strawson claims that when "narrative" can be simply replaced with these expressions, the reference to narrativity seems superfluous.[33] However, the narrative model can resist these charges, because it does not merely claim the narrativity of the self as some exercise in story telling that we can build around these episodes. The narrative model is a theory that conceives of the self as a narrator and makes a claim on the conditions of agency: in this sense then, the appeal to narrativity is not at all superfluous.[34]

---

[30] See Strawson, 2004, for a number of objections to narrative views of the self. Here I discuss the charge of triviality which I believe is the one most relevant for Velleman's model.
[31] Strawson, 2004, p. 73.
[32] We can find such a claim, for example, in MacIntyre, 2007.
[33] Strawson, 2012, p. 76.
[34] In her reply to Strawson, Marya Schechtman has provided a taxonomy of narrative theories, differentiating her own "middle view", which Strawson takes to be wrong, from "weak views" to which Strawson's charge of triviality is addressed and from "strong views", of which Strawson is particularly suspicious and that he actually considers pernicious (see Strawson

Another advantage of the narrative model is that it accommodates other types of acts that might seem prima facie problematic for other authors who also attribute crucial importance to reflective endorsement. For example, narrative coherence does not impose any condition on the wholeheartedness of the agent, as Frankfurt's theory does.[35] As long as the alternatives the agent faces are intelligible to him and can fit within his narrative, his being half- or wholehearted towards his action does not affect the extent of his agency. In fact, as long as ϕ-ing makes sense to the agent, there seems to be no requirement for him to wholeheartedly avow a certain course of action. Doing something only half-heartedly does not in any way mean that I find such a thing less intelligible. Of course the fact that I wholeheartedly prefer some alternative can bear on its intelligibility within my narrative,[36] nevertheless wholeheartedness is not necessary for agency. This also allows for the agent's reasons to not necessarily recommend an action by presenting it as a good or a valuable thing to do: in fact, one can indulge in behaviour that one understands under some negative descriptions, in the light of negative emotions and moods like self-destructiveness, or despair.[37]

Moreover, the role of narrativity promises to be plausibly integrated with our social practices. This is something that Velleman discusses briefly[38] but which can be seen as resulting from the features of narrativity listed above. Narratives being provisional and meaningful opens the way to understand the importance of *social interaction* in their building. Narratives are under an intersubjective influence, inasmuch as they convey a meaning that can be recognized and shared (as well as questioned) by others. Moreover, since the

---

2004, p.447; Schechtman, 2007, pp. 159-161). I cannot present Schechtman's discussion here, but I take it that Velleman's would also qualify as a middle view in her taxonomy and it would then be interesting to see how Strawson's charges apply to his model.

[35] For an insightful criticism of Frankfurt's position and its relation to Velleman's model see "Identification and Identity" in Velleman, 2006. Notice that the narrative model can thus accommodate conflicting motives and changes within the agent's preferences, given that it only requires one's motives to fit within one's narrative in order for them to count as reasons.

[36] And conversely, it could also be the case that when I establish one alternative as more intelligible, the result is that I endorse such a course of action wholeheartedly.

[37] For an extended discussion of despair and negative moods as 'reasons in the light of which I perform my actions' see Velleman, 2000, pp.120-121.

[38] Velleman, 2007b, p.289 and more substantially in his recent discussion in terms of "doables" as social construct, see Velleman, 2013.

provisional character of narratives circumscribes them without isolating them from the web of the agent's other beliefs, others can play a role in our self-understanding because the meaningfulness assigned to certain contexts and scenarios is socially constructed.[39] A major part of what constitutes and shapes your self-understanding is given by the very practice of telling stories to each other.

It might seem that some acts cannot be easily reconciled with the commitments of the narrative model: these are those done "out of character", or that seemingly require some drastic and sudden change in the agent: these acts do not seem to fit into some existing narrative. However, as long as the agent is in the position to make sense of such out of character behaviour or sudden change, he's still able to integrate it within his self-conception. All that Velleman's theory requires for these cases to count as actions is their being intelligible to the agent. Even the most seemingly dramatic conversions can retain INS. This does not mean that one has to explicitly articulate the narrative of such acts in his thoughts. Moreover, the episodic character or narrativity allows actions to be considered in their intelligibility in a specific context, without necessarily affecting one person's character. After all, taking a day off as an impromptu decision might still be intelligible under the circumstances, even for the usually hardworking person, without this entailing any puzzling or dramatic change in the person's character.

Despite these considerable attractions, Velleman's model is unsatisfactory. In the following section I explain why. As in my discussion of Korsgaard, I will not challenge the general idea of a relevant distinction between two forms of activities, some of which seem to be peculiarly human and towards which we seem to have special responsibility.[40] My argument will also assume thesis (ii),

[39] Recently, Velleman has been focusing on how practical deliberation is constrained by socially constructed frameworks. See Velleman, 2013, pp. 23-44.

[40] My argument will also assume thesis (ii), and hence I will not try to examine the view that there are two dimensions of selfhood, one of which can be interpreted in terms of narratives. Moreover, I will not question the rather controversial claim that considers minimal self-reference as insufficient for full agency. I take it that this would raise a different objection to the narrative model, an objection that I do not have the space to address here. Moreover, it seems to me that the rejection of (ii) bears on attributing any special role to the higher dimension of selfhood when it comes to agency has the result of questioning thesis (i) and the

and hence I won't try to provide any support for a challenge to the view that there are two dimensions of selfhood, one of which can be interpreted in terms of narratives. I also will not question the controversial claim that considers minimal self-reference as insufficient for full agency. However, I will show that conceiving of reflective endorsement in terms of narrative coherence results in implausible classifications. This implausible classification results from the fact that the model remains cognitively very demanding and this limits the range of behaviour that can count as action. In arguing for this claim, then, I will question Velleman's attempt to save his model from charges of being overly intellectualistic by relying on narrativity.

2.3. The Limits of Narrativity

2.3.1 Narrativity and Self-deception

So far I have presented the narrative model and I have highlighted how it accommodates some seemingly problematic cases that Korsgaard's model could not account for. However, there is a substantial range of acts that it cannot accommodate, namely acts resulting from self-deception. In the type of cases I have in mind, an agent acts, falsely believing that his motivation to act is *p*. In these cases the agent is not intentionally deceiving himself and his behaviour does not spring from the motives he acknowledges as his. Rather, the causes of his behaviour are opaque to him.[41] Given the kind of narrative

---

very claim that there is a significant distinction between full-blown actions and not. Again, I have no space here to expand these considerations and, as I say in what follows, I aim here at rejecting Velleman's narrative model from within rather than by rejecting all its assumptions *tout court*.

[41] I do not need to examine the large debate involving self-deception and in particular its connection with confabulation here. I will just treat it as the well-known and widespread phenomenon of people having ill-grounded beliefs – where they could and should recognize them as ill-grounded – and performing actions based on them. I do not need to discuss here the distinction between self-deception and confabulation, nor whether or not the subjects are experiencing any tension in self-deception. For discussion see Hirstein, 2006; Mele, 2001. The general phenomenon is sufficient for the present discussion. Velleman, 2006, p. 229, discusses such cases to confirm our striving for self-knowledge. They are examples of failure of the intelligibility drive to direct behaviour: what happens is that the drive then catches up, trying to make sense of what has been done. Nevertheless, for this discussion what is interesting is that the subjects, in conjuring up retroactively a motivationally relevant attitude, deceive

that, according to Velleman, necessarily has to convey the agent's understanding, no cases of self-deception can count as an action and no narrative can be made about them.

The problem with these cases seems to be the same that Velleman claims to identify for Davidson's theory: considering them actions would allow an agent to "act for reasons of which he remains unaware".[42]

My concern here is that the exclusion of self-deception cases would unduly restrict the range of human actions because, in fact, cases of self-deception are widespread; indeed, according to some, they are pervasive.[43] It might be tempting to reply that there is no real problem in excluding such cases from the domain of actions, when other instances of behaviour like slips or some emotional reactions are also excluded. But the problem here is different: the discussion of these cases shows that the theory does not have the resources to establish whether some instance of behaviour is or is not an action. Moreover, excluding them reduces the model's explanatory power, since it would mean that it could not account for a substantial portion of behaviour that is actually one of the main focuses of research in philosophy of action.[44] Cases of self-deception are also harder to exclude than slips or emotional reactions, since in the latter cases, the agent can fail to experience agency or can experience some loss of control.[45] This is not the case in self-deceiving agents.

---

themselves about it and can exhibit the same phenomenology as if they were actually in control.

[42] Velleman, 2007a, p.193.

[43] See Carruthers, 2009, Arpalay, 2003.

[44] See Davidson, 1985 and 1987; Mele, 2001. I do not exclude the possibility that some cases other than self-deception can present problems for the narrative model. The present discussion is aimed at isolating the features of narrativity that generate this problem in these cases and potentially in others that I do not consider.

[45] In general I am not claiming that the experience of agency is necessarily associated with actual agency. I am here stating that experience makes it harder to discard these cases. However, it is important to notice that such an experience might be retained in cases of agents that are, for Velleman, paradigmatic examples of behaviour which is less than fully acted. For example, while arguing with my friend I might have the impression of being in control (even assert that I could stop arguing anytime) while actually being carried away. I do have some grasp of the situation and my behaviour responds to some act-description (proving my point when I have been wronged). However, it fails to be supervised, inasmuch as it lacks coherence with other motives that are still relevant to me. A certain self-conception is controlling my behaviour, but not by means of reinforcing some motives through my understanding of them: rather, by overwhelming and screening others. It does not fulfil the standard of coherence: supervision fails because my understanding fails, thus not inhibiting my behaviour. When

Although Velleman's criterion of conscious control is clear enough to distinguish between actions and mere activities, when applied, it seems to deliver wrong or, at least, counterintuitive results: it is odd that the theory includes as full-blown actions some almost mindless activities, like getting up in the morning or getting a beer from the fridge, but excludes cases of self-deception, which can lead to quite articulated paths of behaviour: both in terms of length and complexity. The agent who bears a subconscious grudge against his parents might genuinely believe that his lies and behaviour are motivated by his desire to be with his friends and, while he fails to recognize his real motives, his behaviour can indeed be rather complex, encompassing possibly quite a wide variety of acts and plans. One could reply that the different actions that are parts of and compose these larger ones can be subsumed under some relevant act description, which seems to at least guide that particular act. So, for example, one can fail to understand his behaviour as an attempt to hurt his parents, but when he organizes to meet up with his friends he can still understand under some description the things one is doing, like 'calling my friends' or 'lying to mum about x or y'. Because the agent retains this understanding, one could argue that his behaviour counts as fully acted. However, I do not think this reply succeeds: the idea of a narrative functioning as a coherence test for endorsement was grounded on the idea that people ordinarily place their behaviour at 'high' or 'comprehensive' levels of description.[46] This integrative self-conception in the light of which the parts of the story are intelligible (for example, conceiving myself as resentful, unsatisfied etc.) is exactly what is missing in cases of self-deception.

I believe that the origin of this problem can be found in Velleman's conception of narrative, which seems to commit the theory to implausible standards of accuracy and correctness about one's own reasons. These demanding criteria show how high the standards of full agency are in the narrative model: so high

---

carried away in this fashion, the agent can retain the impression of control. After the heat of the argument is gone, though, and the balance of his motives is regained, he is left with the one-sidedness of his behaviour. He enacted one character, just not one he can call himself.

[46] Here Velleman refers to the work of Wegner and Vallacher, 1986, see Velleman, 2006, p. 249.

that they seem implausible, as they do not allow any slack between our conscious and reportable intentions and our true motives.

One could try to relax these demanding standards for self-understanding by claiming that the agent does not need to be *right* about his motives: as long as something makes sense to me and I can provide myself with at least *some* reasons, the intelligibility drive can reinforce and motivate me to perform the action my reasons recommend. This reply is based on the idea that the self-deceiving agent, conjuring up an interpretation of his behaviour, would satisfy both the intelligibility and the self-consistency drive, which Velleman claimed were the psychological elements at work in actions. However, this reply is not satisfactory. What it assumes is that self-consistency is some mechanism merely sensitive to the correctness of the behaviour it contributes to produce, rather than to the correctness of the self-representation it implements. And that this is sufficient for agency. However, in this picture, when the intelligibility drive conjures up some interpretation something seems to go amiss: it seems to compensate for some failure of one's understanding. In the cases considered here the action is not guided by a correct self-representation and one's self-governance is in fact illusory. The mismatch between my real motives and my purported reasons risks making Velleman's picture collapse into Dennett's, which strictly separated the agent's fictional (narrative) reasons and the actual causes of his behaviour. Velleman's conception of the self as a narrator imposes a requirement of correctness for which the possibility of limitations of self-knowledge seems problematic. Moreover, once we assume the actual limitations of self-knowledge, one's authority in establishing that one's self-understanding is genuinely effective is also limited. As a result, the theory gives us no sure grounds to determine whether some specific instance of behaviour is to count as action or not and whether one's narrative plays a role both "fictive and factual"[47].

I believe that the notion of narrativity adopted by the model commits the theory to these implausible standards of self-understanding. This is

[47] Ibid., p. 221.

particularly evident when we look at the way narrativity interprets the conscious control that is required for agency in Velleman's view. To understand this I will examine more in detail the idea of an *Implicit Narrative Structure* (INS). My discussion here focuses on the term 'implicit', which is ambiguous, for it can be interpreted in two importantly different ways.

Under one interpretation 'implicit' could mean 'not-explicit but readily accessible'. If so, the narrative structure would provide reasons that are not the focus of my attention, but that are readily accessible to me. Even in cases in which I acted seemingly automatically I could, if asked, come up on the spot with rather complex and articulated stories, which were highly unlikely to have been in my conscious thoughts whilst acting. A potentially accessible narrative structure can therefore fulfil the criteria of agency. I know what action F is, under a certain act-description, and I am in control of it, as well as of my motives. Indeed, we can see how an implicit narrative structure could underlie even some expressive reactions, allowing their interpretation as full actions. INS explains how my understanding of the situation elicits a certain behaviour, which counts as supervised and controlled.

But a different interpretation 'implicit' is possible, where it could mean 'not-explicit and not-readily accessible'. So interpreted, the implicit narrative structure that coordinates my action would remain hidden to me. My dissatisfaction in my current relationship might lead me to engage in certain courses of action without my acknowledging that this is my motive and even my denying it, or identifying a different one. This idea introduces an *unconscious* narrative structure (INS*),[48] in virtue of which you might be fulfilling a role you do not understand. It seems that cases of self-deceiving acts involve INS*. We have seen above how Velleman describes the case of a child about to lie to his parents.[49] We can convert this case of INS into one of INS* if we imagine that the child is not able to tell a story about his

[48] It is not necessary, in order to make this objection, to claim that this narrative structure INS* is absolutely inaccessible: Velleman's own example of arguing with a friend provides us with an example of behaviour that fails to exhibit conscious control *during* the performance, but whose narrative structure can be retrieved after some time.
[49] Velleman, 2006, p. 251.

willingness to tell a lie. He might be driven by some subconscious grudge, and in doing so he might be fulfilling a narrative, only unconsciously, not led by his awareness of any self-conception. If so, he could not refer to it as a rationale for his actions: he would be failing to acknowledge his subconscious motives and would be avowing reasons that are not in fact moving him. Behaviour with an INS* is guided by *some* understanding of the situation: an understanding that leads one to choose and select certain courses of action and not others. But this happens, regardless of one's access to the actual guiding self-representation. If INS* could be considered a genuine form of narrative, then, despite the appearance, the cases of these self-deceiving agents would count as actions after all.

However, this solution is not available to Velleman. INS* cannot be considered a genuine form of narrative because it does not have all of the features attributed to narrativity. There seems to be no problem for the suggestions that an INS* has an *episodic* character or that it entails *vagueness*. Moreover, an INS* certainly assumes that the underlying understanding of the situation selects certain responses and not others. In this sense there can be an unconscious description that triggers the agent's responses: although unexpressed, and not recognized explicitly, the *meaning* of a certain situation can shape the agent's understanding. So, one's friend, acting out his subconscious resentment in an argument, might follow specific paths and choose actions or words he "knows" to be more hurtful without acknowledging that he is doing so. Given these characteristics, INS* seem to have at least three of the four features of narrativity: namely, meaningfulness, vagueness and its episodic character. However, it is impossible for INS* to account for the *authorship* criterion: INS* is certainly productive, for the behaviour it produces fulfils a narrative. The problem is that it is not productive in the right way. As we saw above, a necessary feature of authorship is that the narrator takes himself to be the author of his story. This means that the kind of narrator conceived by Velleman is almost an omniscient one: this does not mean that he knows his whole life story in advance (this is not even an issue, given the episodic character of narratives), but rather that, within the small narratives he

authors, he knows all the details that are necessary for selecting a certain step within the narrative.[50] In fact, access to those reasons is required if he is to perform the relevant behaviour *in light* of those reasons. INS* seems to lack the feature that makes a narrative model attractive for a reflective endorsement theory: a narrative self-conception alludes necessarily to a self-representation that has to be accessible. Access is hence required both by agency and narrativity, and it seems built in this conception of narrative. It is this access that shows how the metaphor of narration could explain the connection between self-consciousness and self-governance, and it is required for Velleman's claim that just being motivated by some desire is not equivalent to being guided by reasons.[51] But it is this access that is lacking in cases of INS*.

In fact some of Velleman's own examples of 'less than full-blown' actions can be understood as instances of INS*. One, which has already been mentioned,[52] is the case of meeting up with a friend to solve some disagreement. During the meeting I get progressively angrier and eventually end up arguing. Only later do I realize that 'accumulated grievances had crystallized in my mind, during the weeks before the meeting, into a resolution to sever our friendship'.[53] In these cases I can believe that the decision, though genuinely motivated by my desires, was thereby "induced in me but not formed by me".[54] Behaviour guided by INS* cannot count as fully acted and INS* cannot be a genuine form of narrative according to the model. And this despite the fact that, to return to the example above, telling the lie can be seen as perfectly coherent with my being resentful, and my argument as the perfectly fitting conclusion of my already strained friendship.

So far I have argued that Velleman's conception of narrativity cannot accommodate cases of self-deceiving agents. As a result of this, it seems that the appeal to narratives cannot really serve to address charges of his model

[50] Notice here that the third criterion of narrativity confines the omniscience of the narrator to small episodes; so the fourth criterion of vagueness allows for not setting too high a standard for specificity. Here all the discussion requires is that the agent is at least to be seen as an omniscient narrator in regard to the reasons that motivate his actions.
[51] Velleman, 2004b, p. 279.
[52] See above my notes 45 and 48.
[53] Velleman, 2000, p. 126.
[54] Ibid., p. 127.

being overly intellectualistic and cognitively too demanding.[55] In particular, our discussion of INS* has pointed at the crucial role played by the feature of 'authorship' in committing the theory to demanding standards of self-understanding. In my next section, I will show how this very same feature creates further problems for Velleman's model.

2.3.2 Narrativity and supervision

I now turn to my second point against Velleman, which is that it is unclear that the metaphor of narration can succeed in accounting for agency, once we have – through (iii) – ascribed two distinct functional roles to practical reasoning: one as deliberator and one as supervisor. I will question the compatibility of the idea of conceiving the agent as a narrator with the claim that there are two kinds of reflection.

As we have seen Velleman specified the processes through which practical reasoning takes place in terms of perceptual and deliberative reflection, where the former can be seen as the widespread phenomenon at work in practical reasoning and the latter as only ancillary to it. The narrative module, if constitutive of agency, must be able to work through both of these processes: the self as narrator needs to encompass both the role of the deliberator and that of the supervisor.

Firstly, I want to highlight a distinction between deliberative and perceptual reflection that so far has been only implicit: a distinction between *active* and *passive* in this specific context. I define deliberative reflection as active since the deliberator actively intervenes through his decision in leading his behaviour some way or another. On the other hand, when it comes to perceptual reflection, the agent need not intervene: behaviour, as Velleman phrases it, is 'under way' and as long as it 'makes sense' the agent need not play an active role in directing it. In this sense, the supervisor is passive. We could say that it is theoretically active, but practically inactive, since it is a system of monitoring behaviour that is unfolding without intervention.

[55] Velleman, 2007b, p. 284.

As pointed out above when introducing the distinction between deliberative and perceptual reflection, there seems to be a variety of ways in which the passenger can interact with the driver. He can sit quietly as long as he agrees on the route; he can suggest a different turn or put up a team effort and direct the driver according to the map, he can nod and say, "I told you so" when something goes wrong; he can, eventually, give up on mere supervision and decisively take the steering wheel to oblige the driver to change direction. The metaphor suggests different degrees of passivity and participation in behaviour. It seems that the monitoring can take place as simple acknowledgement and progressively allow the supervisor to play a more active role. For the sake of argument, I will consider the case of a minimal degree of participation. Establishing that there are different degrees of passivity or activity within perceptual reflection is not an objection to my choice of case, since the supervisor is passive in at least some relevant conditions.

Indeed, the very idea of perceptual reflection was introduced to make sense of the intuition that automaticity, not deliberative thinking, is the rule and not the exception. In this non-active position, the supervisor shapes the intelligence of the sub-personal driver. As described above, in its supervisory role, practical reasoning stays in the background both of action and thought. Identifying the agent with the supervising intelligence means that it was there all along, just "letting" the body move. To sum up, a core characteristic of Velleman's supervisor is its being passive, even if there are degrees and active dimensions within perceptual reflection.

Given this distinction between active and passive supervision, I want to argue that a core feature of narration is its being a form of *activity*. *Pace* Velleman, narration is not a theoretical stance that the author has towards some events. It implies the practical task of producing a story, engaging in some mental composition of a story. As an author, the narrator is active: he is not told, but rather he tells the story. There would be no story at all without at least this intervention. Now, one can claim that in supervising, as in narrating, one can be considered the author of the results the supervision leads to. Narration,

however, is an activity of a different sort than supervision is: one that entails an active intervention of the narrator in the composition of his story. This is not to say that a narrative needs to contain all the details, since we have already seen that narratives allow a certain degree of vagueness.[56] Nevertheless, if a scenario or a framework of meaning needs to be built, the narrator is active in its composition. If the narrator was to assume a merely supervisory position with respect to his story, (letting the story "unfold", so to speak), there would be no story at all. It is really this practical dimension of narrativity that allows the metaphor to present the agent as an author, grounding Velleman's conception of action.

Is saying that the narrator is always active merely an *ad hoc* stipulation? One could claim that these specifications associated with the metaphor of narration are not essential to Velleman's picture. However, I think that the discussion above served to show how the metaphor of narration carries important implications in terms of the functions ascribed to the self. The picture resulting from considering narrativity as constitutive of full agency gave us a view of practical reasoning as expressing one's fundamental drive to self-understanding: you need to find yourself intelligible as the character you take yourself to be. The way you succeed at satisfying this need is by giving expression to your thoughts, feelings and personality through the things you do:

> Here the verb *giving expression* describes a substantive activity. In ordinary actions you do not *let* your thoughts, feelings and personality *come out* through your body, as if they were seeping through your pores. Instead of *acting out* in this manner, you perform intelligible actions by selecting a coherent subset of your thoughts and feelings, composing a coherent expression of them and thus producing words and movements with a unified message.[57]

Velleman's view of action requires this characteristic of narration as part of the authorship criterion, in order to resist Dennett's conception of the self as purely fictional and not to cast the agent in the merely contemplative position of a commentator. As we have seen, Velleman's idea is that the agent's self-

[56] See above, p. 50.
[57] Velleman, 2008, p. 422.

conception feeds back into behaviour and it does so by means of actual agential control.

There seem to be a clear tension, therefore, between the characteristics Velleman says are necessary for the authorship attributed to the self as a narrator, and the role it is supposed to play in perceptual reflection. If the narrator is always active, as I argued above, then the supervisor is not a narrator and narrativity is confined to deliberative reflection. In short, Velleman needs to explain how perceptual reflection can be an actual instance of practical reasoning, given the authorship requirement.

One possible solution to this underlying tension would be retain the requirements of authorship imposed by the narrative model and deny (iii) – that the things we do through perceptual reflection would not count as actions at all. However, we have seen that (iii) cannot be abandoned without important consequences for Velleman's model because it is precisely what allowed it to encompass a wider range of actions than a model like Korsgaaard's, which conceived reflection as deliberation. This solution would therefore have the consequence that the narrative model would really cover a very limited range of cases. I have already explained in my previous chapter the reasons why limiting the domain of actions to cases of deliberation is clearly excessively restrictive. Another possible solution would be to not consider narrativity as constitutive of agency. This could lead to confining narrativity to cases of deliberation or to making it crucial for self-understanding rather than for agency. This would indeed solve this problem but would create another. For now Velleman would have to define another kind of coherence test, or another kind of test altogether, to provide the endorsement necessary for something to count as an action.

To sum up, in my first objection I showed that Velleman's model necessarily excludes cases of self-deception and argued that this is due to the high (and implausible) standards of self-understanding implicit in his conception of the

authorship criterion. In the objection just discussed I showed how the metaphor of narration is problematic if we want to account for what Velleman thinks is the merely supervisory role that practical reasoning sometimes has in our actions. Again, I stressed how the authorship criterion forces this conclusion. Since, as we have seen, the source of the trouble with Velleman's conception of narrativity is the feature of 'authorship', which identifies a narrator's conscious control with agential control, one could try to answer both concerns I raised by introducing a different conception of narrativity: one that redefines or excludes this authorship criterion. In the final section of this chapter I will briefly discuss this proposal and show why the authorship criterion is not dispensable in Velleman's project.

2.3.3 Narrativity and authorship

From what we have seen so far, for Velleman the narrator embodies the idea of a 'controlling consciousness' because his knowledge produces the story he enacts. Consciousness is also a requirement for narration inasmuch as one cannot narrate what one does not know. Like an agent, a narrator needs to check that his story coheres with the intended meaning, thus enabling supervisory as well as directory functions. If one were to reconceive of narrativity by either reformulating or eliminating the authorship feature, one would also have to reconsider the relation between the control a person has as an agent and as a narrator.

According to Velleman you exert control through consciousness of what you are doing. It could seem, then, that one could try to retain this picture while giving up on authorship or loosening up its standards, thereby untying the conscious control attributed to narrativity from causal control. A consequence of this is that the kind of resulting narrative control, if not understood in causal terms, seems to amount to a kind of understanding. If so, we confine narrativity to narrative thinking or narrative explanation. This suggestion might seem promising because it reflects the idea that narrative understanding is not just a mere causal-psychological explanation. Moreover, this would be more fitting with the theoretically active but practically inactive role assigned

to certain forms of supervision. However, this way of reformulating narrativity requires us to reconsider the relationship between narrativity and causality. This is because it separates the conative and cognitive components that Velleman strived to connect,[58] with consequences for the unity of agency. This means that the narrator and the protagonist would come apart. In so doing one would redefine agential control, by differentiating a causal component and an understanding component, allowing narrativity to play a crucial part only in the second one of them. Since the directive consciousness that Velleman requires for actions is not mere understanding, one would have to provide some further story for the mechanism that leads to the production of behaviour, and without that it would seem that the narrative model falls short in accounting for the conditions of agency.

Alternatively, one could reconsider the role played by reflective endorsement. If one were to hold on to the idea that even in this reading action is behaviour one has reflectively endorsed, the resulting picture could be one where action is behaviour we have reflectively endorsed regardless of how it came about. In this sense, phenomena like self-deception or confabulation would cease to be problematic. Action would be behaviour we can commit to, that we can explain: for something to count as action I would need to be in the position to talk about it and justify it in the light of my narrative understanding, regardless of its causes and my awareness or accurate understanding of them. After all, one could say, we were after the specifics of human agency. The difference between humans and animals was traced back to the possibility of an objective self-representation and such a feature is independent in principle from its role in producing behaviour, or from an accurate understanding of it. In fact, self-deception is surely a distinctively human phenomenon.

Unfortunately, however, I don't think that this solution is viable for Velleman's project for two reasons. Firstly, because in this reading the notion of conscious control that according to Velleman is necessary for agency remains unclear:

[58] See for example, Velleman, 2000, p. 25 where Velleman understands the choices leading to actions as practical cognitions: they have the cognitive direction of fit and the associated aim of being true and the practical direction of guidance.

this proposal seems to rely exclusively on one's understanding and such an understanding seems to bear no necessary connection to one's experience of agency. Secondly, even if one succeeded in providing an explanation of the connection between narrative understanding and agentive experience[59], this view of the role of narrativity falls short of securing the general goals of reflective endorsement theories of agency. We have seen how what these theories want to show is that self-governance is the distinctive feature of human actions and that reflective endorsement is the actual condition to choose "our own causality"[60], to use an expression we already found in Korsgaard. The interpretation of narrativity I am now discussing clashes with this position. This is because in this new picture, apart from the kind of story that the agent could provide, behaviour would not be caused in any different way in the human case and in the animal case. In the original formulation of Velleman's narrative model that I presented in this chapter the relationship between self-consciousness and self-governance was secured by the connection between narrativity and the cognitive motives which govern human motivation. In the present formulation of narrative control, this connection is severed. We can see this if we consider, for example, how this view would account for the fact that a subject could understand his behaviour and explain his reasons while in fact he did not actually bring it about.[61] In these cases of misattribution it would be wrong to think that the agent actually performed the action he attributes to himself. Misattributions of this sort show that reflective endorsement which results from narrative coherence could occur entirely post hoc. It is unclear how post hoc understanding could function to determine the motivation that leads to action so that the behaviour counts as self-governed. This conception of narrativity would then not serve in accomplishing the goals originally set by the reflective endorsement approach. In fact, if, in trying to provide an alternative conception of narrativity, one were to follow a strategy to differentiate between a narrator's control and

[59] *Top-down* theories of agentive awareness seem in fact to argue that the sense of agency emerges from this kind of high-level mechanism subsuming of one's action under comprehensive explanations which many authors describe in narrative terms. See Roser and Gazzaniga, 2006, and Stephens and Graham, 2000, p. 161.
[60] Korsgaard, 2009, p. 108.
[61] Pacherie, 2010, p. 454.

causal control, one would still need to provide some story about the relationship between self-consciousness and self-governance.

Alternatively one would have to limit the role of narrativity in ways that radically downplay the role Velleman originally ascribed to it. For example, one could conceive of narrativity as a general characteristic of many, if not all, our actions, but not a criterion for agency. One could still sometimes happen to enact and fulfil one's self-narratives, but this would not be necessary for agency. It might even be the case that narrativity and agency could pick out the same set of doings, but narratives would figure in the picture in purely non-causal terms. Moreover, giving up on a direct connection between narrativity and causal control exposes this conception of narrativity to Strawson's charge of triviality that I briefly presented earlier. We would need to reconsider the extent to which our alternative conception of narrativity is not trivial in Strawson's terms. As we saw above, Velleman's narrative model resisted this charge of triviality by its being a theory of the self as a narrator, rather than of the narrativity of the self. A solution that, once we untie narrativity and causality, does not seem viable anymore.

2.4. Conclusions

In this chapter I questioned the idea that narrativity can actually serve as a criterion for agency focusing on Velleman's picture. Despite the fact that narrativity seems connected with our reason-giving practices and that Velleman's model promises advantages over Korsgaard's, I argued that his model, like Korsgaard's, fails to provide us a satisfactory understanding of agency. I argued that the exclusion of cases of self-deception is problematic because it shows that the narrative model is too restrictive, precisely because it is committed to too demanding standards of self-understanding. I also pointed at the difficulty of integrating the idea of two kinds of reflection, deliberative and perceptual, within the metaphor of narration. I claimed that these issues arise from Velleman's conception of narrativity, and I considered how altering

such a conception, so that, for example, one could separate narrative, as a form of conscious control, and causal control, would avoid some of these problems but it would also require us to reconsider the model altogether.

In my discussion so far I have left open the possibility that narratives could play some other significant role in agency, and I also have not challenged Velleman's overall conception of practical reasoning. However, I hope to have shown that Velleman's notion of narrative cannot address the charge that his model is overly intellectualistic and cognitively too demanding. In the following chapters I will argue that these shortcomings are in fact grounded in two problematic assumptions which are also present in Korsgaard's picture, namely a cognitively very sophisticated conception of self-consciousness and the obscure notion of control necessary for human agency, that both models rely on.

Chapter 3

The Nature of the Self: Metaphysical Commitments and Representational Conditions.

In the last two chapters I examined Korsgaard's and Velleman's models of agency, in which the concept of reflective endorsement plays a constitutive role. Much of my discussion so far has revolved around the account of reflection these models provide, because this is a core concept that any account of reflective endorsement needs to make clear. I found their views problematic on two counts. First, they place overly intellectualistic constraints on their notion of agency, and second, they unduly restrict the range of behaviour that can count as action. In this and the next chapter I will highlight how these difficulties arise from the answers to the other two questions I presented: one about the relationship between self-consciousness and self-governance; and the other one about the conception of the 'self' entailed by their notion of self-consciousness. As a result, I will question the view, shared by these two models, that agency is behaviour guided by a representation of a certain kind, linked to one's self-conception and one's self-understanding. I believe that there are two problems with this definition: one is the kind of representation necessary for self-governance and which allows the control over our motives. As we shall see, the way both models conceive this self-representation is the root of the difficulties examined in chapters 1 and 2. The second one is the conception of guidance and control employed by both views of "self-governance", which I will claim is obscure and underdeveloped. I will discuss the first issue in this chapter and the second one in the next.

My present discussion will focus on the notion of "self". Both views consider the "self" as a distinctive product of the human capacity for self-consciousness, and in both models it plays an essential role in defining the threshold of human agency. I will try to show how ultimately both views focus on an aspect of self-consciousness that requires a high level of intellectual sophistication. Because of this, they fail to fully capture for the specificity of human agency, for which more minimal forms of self-awareness are sufficient. This analysis of the notion of the self will allow me to consider the idea of a 'narrative self' in more depth than I have done so far. There is much debate around this notion[1] and we find substantial heterogeneity about how it is understood in different contexts. I will then take the present discussion as an opportunity to discuss Velleman's account within the framework of other conceptions[2] of the narrative self. I will structure my discussion of the self around two questions. Firstly (§3.1), what is the 'narrative self'? In order to answer this question I will explore the metaphysical commitments of this conception of the self and connect these considerations to the role the self plays in defining the specifics of human agency. Secondly (§3.2), is the narrative self, in Velleman's model, an illusion? I will argue that answering this second question is of vital importance for Velleman's project and I will show that Velleman's theory does not in fact settle the issue.

In §3.3 I connect the results of this discussion to Korsgaard's model and question the plausibility of what I presented in the previous chapter as the 'two levels of selfhood' thesis, which Velleman and Korsgaard share. I will argue that the kind of self-representation that these models claim to be distinctive of human subjectivity and human agency restricts self-consciousness to an excessively high level of cognitive complexity. As a result, their models of agency are bound to present the difficulties I have highlighted in previous chapters.

[1] Views advocating for or against the idea of a narrative self can be found, among others, in Schetchman, 1996; Hutto, 2007; Dennett, 1991; Strawson, 2004.
[2] In particular, I will confront Velleman's conception with Dennett's, but also with Goldie's and Schechtman's.

3.1. What is the narrative self?

We can understand the question about what the narrative self is as concerning the metaphysics underlying the narrative account. What *sort of thing am I* according to narrative theories of the self? What are the metaphysical commitments of the narrative view? Am I an immaterial substance? A bundle of thoughts? Or something else entirely?

The traditional Cartesian account would hold that I am an immaterial thinking substance. Hume suggested that each of us is a 'bundle of perceptions'. Among contemporary approaches, according to animalism you are that human animal located where you are,[3] while according to psychological approaches,[4] although you are made of the same matter as a certain animal, you and the animal are different things with different persistence conditions: some psychological relation is taken to be necessary and/or sufficient for you to persist. Alternatively, some have described the self as inner, unidentifiable and ineffable.[5] It is important that all these answers place the self in the fundamental ontological category of 'things', or 'objects', rather than understanding it as a 'property' or an 'event'.[6] Only in virtue of being some sort of thing, or object, can the self function as reference of the word 'I' and be the subject of my attributions. Both animalists and proponents of psychological approaches conceive the self as an object defined by its fundamental properties, while they differ in their accounts of such properties: psychological for the latter and biological for the former. This means that the self can function as a subject of predication and as a bearer of attributes, and is not itself predicable. Most importantly, the self being a target of reference allows different theories to present subjects as entitled to some claims of the form "I am x". This is the case even for views that do not retain much appeal nowadays like Cartesian dualism: the dualist can claim that essentially, "I am an immaterial thinking substance", while it would be false, according to this

[3] Olson, 1997; Snowdon, 1995.
[4] Shoemaker, 1984; Parfit, 1984.
[5] Harré, 1984.
[6] I remain neutral about the definition of what events, properties or processes are and their relations. It is sufficient for me to oppose them to the category of objects, entities or things also loosely defined.

account, to say that "I am an animal". In this view, other ordinary self-attributions will need to be adequately rephrased: so "I am 64kg" will in fact be rephrased as "I have a body, and my body weighs 64kg". Such claims need not to be more mysterious than my stating that "I don't think I fit in that space", while I am parking my car: by which of course I mean that it is my car that would not fit in that particular spot.

To sum up, in answering the question "What am I?", selves are conceived as subjects of action,[7] bearers of properties, and reference for true or false attributions.

How does the idea of a narrative self fit within this picture? Is the narrative theory of the self an alternative to these accounts? Is the narrative theory committed to any of them?

Much of the debate around the narrative self has been developed from Dennett's original formulation,[8] which was grounded in the conclusions he drew from cognitive psychology in his *Consciousness Explained*.[9] Dennett's is the immediate reference for Velleman's work, and I shall now examine the former's theory at more length than I have done so far.

3.1.1 Dennett's view of the narrative self

Consider the following passages in which Dennett develops the idea of a narrative self:

> A self…is an abstraction defined by the myriads of attributions and interpretations (including self-attributions and self-interpretations) that have composed the biography of the living body whose center of narrative gravity it is. As such it plays a singularly important role in the ongoing cognitive economy of that living body, because, of all the things in the

[7] This echoes Leibniz's definition of substance as "that which acts", see Leibniz, 1989, p. 207. When I conceive what I am along these lines we can make sense of the fact that for example, I can truthfully utter "I am writing a chapter of my thesis".
[8] As opposed to continental and phenomenological traditions like Ricoeur's conception of the narrative self we find in Ricoeur, 1988.
[9] For example, Gazzaniga, 1985.

> environment an active body must make mental models of, none is more crucial than the model the agent has of itself.[10]
>
> On the view of selves - or persons - emerging here…selves are not independently existing soul-pearls, but artefacts of the social processes that create us, and, like other such artefacts, subject to sudden shifts of status.[11]
>
> Thus do we build a defining story about ourselves, organized around a sort of basic blip of self-representation. The blip isn't a self, of course; it's a representation of a self…What makes one blip the me-blip and another blip just a he- or she- or it-blip is not what it looks like but what it is used for. It just gathers and organizes the information on the topic of me in the same way other structures in my brain keep track of information on Boston, or Reagan, or ice cream.[12]
>
> And where is the thing your self-representation is about? It is wherever you are. And what is this thing? It's nothing more and nothing less than your center of narrative gravity.
>
> But don't I exist?
>
> Of course you do. There you are, sitting in the chair, reading my book and raising challenges. And curiously enough your current embodiment, though a necessary precondition for your creation, is not necessarily a requirement for your existence to be prolonged indefinitely…If you think of your self as a center of narrative gravity, your existence depends on the persistence of that narrative, which could theoretically survive indefinitely many switches of medium…If what you are is that organization of information that has structured your body's control system … then you could in principle survive the death of your body as intact as a program can survive the distraction of the computer on which it was created and first run.[13]

In these passages Dennett seems to make many heterogeneous claims about the nature of the self. On the one hand he seems to have a phenomenal understanding of the self, as the experience of oneself as the conscious subject; on the other hand he seems to make some metaphysical claims. He suggests that the self is an abstract artefact of social processes, that the self is a fictional abstraction with a distinctive practical role, and also a function of the organism, a biological product, an organization of information. At the same time, he also urges that selves are also persons, and seem connected with the subjects we refer to when we raise questions of personal identity and survival. As it is, the view is far from clear. I will try to clarify Dennett's position in

[10] Dennett,1991, p.426-427.
[11] Ibid., 423.
[12] Ibid., 429.
[13] Ibid., 430.

light of the considerations I drew above about the self being subject of action, bearer of properties and reference for true or false attributions.

One strand of Dennett's thought presents the self as an illusion conjured up by the organism. Is it possible to consider this illusion as the subject of actions? According to Dennett, human organisms have particular properties that allow them to function within the requirements of their complex environment: among these, the self's function is to enhance our behaviour "by the illusion of greater unity". It is clear that, in Dennett's view, it is the organism that is in fact doing all the causal work and the self cannot be credited with actual causal powers. In fact, the self is just a function of the organism which organizes information around a certain unity of narrative, by creating the illusion of a "central meaner and controller".

Is the self a bearer of properties? In Dennett's view, *it just appears to be so*: the unified subject of one's experience is, in fact, just an illusion produced by my brain. It might seem that, if the narrative self is a tool, a process employed by the organism to play some function, it needs a user, which could be the brain and which could be then thought as my real self. Now, if the brain were the subject of experience, in asserting that I exist, as such a subject of experience, it would be true that in fact I am my brain. Some surprising consequences seem to follow from this. For example, does my brain own these clothes, for example? Dennett claims that this is not the case and rejects the idea that my *real* self could be the brain. He thinks that your current embodiment, although necessary for your coming to be, is not "necessarily a requirement for your existence." Dennett not only aims at clearing the stage of the Cartesian theatre of all ghosts and declare the very stage as not existent, but, in parallel to his denial of Cartesianism, he also opposes the kind of reductionism that would identify the self with any part of the organism.

Dennett advocates a form of fictionalism, that involves some kind of elimination: there are no such things as selves, as "central meaners and controllers", rather, we should actually understand that selves are fictional abstractions naturally originating from our linguistic capacities. These fictional

abstractions enable us to function in our environment and are necessary for our social interactions. According to Dennett, then, our everyday and folk-psychological understanding of selves is the result of some kind of error, one that "is difficult, but not impossible"[14] to get rid of. Dennett claims that we can understand how the error came to be and what its function is, and he also claims that no substantial revision of our vocabulary is needed. You are still entitled to say that you own your clothes, like ice cream, and that you disagree with some current local policy.[15] This claim raises the question in what way our ordinary self-attributions are preserved if we identify selves with fictional abstractions. Fictional objects here seem to be understood as abstract artefacts: abstract inasmuch as they bear no spatial properties, and artefacts since they are the result of human activity, and depend on an author narrating them.[16] Intuitively, the "self" seems dependent on the subject's psychology of which it is a function. If it wasn't dependent on the subject's psychology, it could seem that *my* self could be instantiated and authored by different subjects.

The idea that selves are abstract artefacts has repercussions for personal identity and survival, which would not be much different from those we attribute to fictional characters. Fictional characters are individuated by the act that brings them into existence, but they are necessarily not complete. We do not know in what position Achilles was sitting after Priamus left his tent, or whether Holden Caufield left the park whistling on the way home. But these gaps could be completed, and not necessarily by one author only. For example, much debate has revolved around the existence or not of Homer, where an alternative theory suggests that the Iliad is the product of multiple authors. Regardless of there being an historical Homer or not, Helen would still be the same character. Similarly, Dennett notices that it is possible for one organism to have multiple selves (as in the case of patients affected by Multiple

[14] Ibid., 424.
[15] "The narrative self is yet another abstraction, not a thing in the brain, but still a remarkably robust and almost tangible attractor of properties, the "owner of record" of whatever items and features were lying about unclaimed. Who owns your car? You do. Who owns your clothes? You do. Then who owns your body? You do! When you say 'This is my body'. You certainly aren't take as saying 'This body owns itself'." (ibid., 418).
[16] Sainsbury, 2010, pp. 91-92.

Personality Disorder) as well as, in principle, one self for multiple subjects.[17] This is the case for Dennett because the self is individuated by its function, which is rooted in the pressure of our social environment.[18] Moreover, as already noted, Dennett holds that survival could in principle not be affected by switches of medium and material constituents.

A last aspect that needs to be considered is whether or not the self, as conceived by Dennett, allows claims I can make about *myself* to be true. Dennett characterises the self as a function of an organism, while also stating that the self is a fictional object or artefact. These seem to be two rather different claims. A function is something that we, organisms, have, rather than are, and it can hardly work as a subject. Attributing anything we ordinarily do to a brain function would seem no less problematic and would require an even more dramatic change in our normal use of language than claiming that "I am this brain". After all, my brain is a concrete object, and as such a bearer of properties. If the self is a function of the organism, it seems just to be the wrong ontological category for me to be in. This makes it very hard to understand any claim I can make about myself, since functions and processes are the wrong candidates to be subjects.

On the other hand, as a fictional object, for Dennett the narrative self is a "remarkably robust and almost tangible attractor of properties":[19] the fact that we can make several attributions referring to fictional characters seems obvious. However, it is not obvious that we attribute properties to fictional characters in the same way, with the same attitude, that we normally reserve

[17] "The convictions that there cannot be quasi-selves or sort-of selves, and that, moreover there must be a whole number of selves associated with one body - and it better be the number one! - are not self-evident…MPD challenges these presumptions from one side, but we can also imagine a challenge from the other side: two or more bodies sharing a single self! There may actually be such a case, in York, England: the Chaplin twins, Great and Freda (Time, April 6, 1981). These identical twins, now in their forties and living together in a hostel, seem to act as one, finishing each other's sentences with ease or speaking in unison…some who have dealt with them suggest that the natural and effective tactic that suggested itself was to consider them more of a her." (Dennett, 1991, p. 422).

[18] These considerations also highlight how Dennett's narratives are built: many details of my self-narratives might not literally exist until our social discourse demands public explanation of me. The parallelism with fictional characters can hence be brought forward by noticing that, since the minimal conditions to possess such narratives require linguistic and conceptual mastery, selves are possible only for creatures with such abilities.

[19] Ibid. p. 418

for real people in our ordinary discourse. It is essential to notice that in making predications about fictional characters we are assuming that our utterances are made in a fictional spirit and set a context where they can be claimed true so that for example we say: "According to Shakespeare's story, Titus cut his hand". If we were to take seriously Dennett's claim that no substantial revision of our language is needed, it would seem that our talk about selves can be interpreted as also tacitly within a fictional context that prevents issues about its truth to arise. However, this is not the attitude that we seem to have towards these claims about ourselves and neither does it seem to be the attitude that Dennett claims we should have.[20]

In fact, Dennett's eliminativism seems to require much more revision of our ordinary discourse about ourselves than he supposes, if we are to be entitled to "I" statements in a purely fictional way, in the same way we can attribute properties to fictional and abstract objects. How can the comparison with fictional characters help with the difficulties in our ordinary language attribution? The way in which abstract objects like fictional characters figure in what we say and think, along with the possibility of assigning truth values to our assertions, is subject to much debate and I will not be able to do it justice here. One influential proposal to deal with predication about abstract artefacts distinguishes between two ways in which properties can be related to them: abstract artefacts can either *exemplify* properties, which the entity really possesses, or they can *encode* properties, which are ascribed to the story. So, for example, when we say 'Holmes smoked a pipe', Holmes would be encoding the property of smoking a pipe but not exemplifying the property.[21] What Holmes exemplifies is 'being an abstract object', or 'encoding the property of smoking a

[20] see ibid. I do not think that we should interpret Dennett's as a form of revolutionary fictionalism, and the ontological thesis that identifies selves with fictional entities does not run together with a linguistic thesis. He is not just suggesting that we need not dismiss the claims we make in our ordinary discourse because it is worth keeping for various purposes, most importantly, since the self, as a fictional character of our narratives, is a natural product of our linguistic abilities and is fundamental for our social practices. He literally sees no problem in our ordinary usage of the language. To him this just seems to mean something different from what we had thought when we assumed soul pearls and homunculi.

[21] I consider this distinction as presented in Sainsbury, 2010. Sainsbury credits Zalta (1983) with this view and considers how this is essentially the same distinction we find in van Imwagen (2003) between having versus holding. See Sainsbury, 2010, pp. 93, 224n.

pipe'.[22] Can we apply this proposal in order to make sense of fictionalist claims about the "self"? I do not think we can. Proponents of this view argue that, if we do not make this distinction between encoding and exemplifying, our statements about fictional characters present an ambiguity of predication that results in controversies about their truth value. The problem with Dennett's fictionalism about the self is different because claims about my "self" retain a fundamental ambiguity regarding the reference of 'I' rather than regarding its predicates. How are we to understand a proposition like "I play the violin"? Can we say that I, as a fictional character, encode but do not exemplify the property of playing the violin? Next to an ambiguity of predication there is an ambiguity in the identification of the subject: I, as the fictional narrative, encode the property, while it is possible that I, as this organism whose functional property the narrative is, exemplify playing the violin. It is not in fact possible for "me" as a fictional abstraction to exemplify playing the violin.

In light of this, it is not clear that fictional discourse can help fixing a reference that would maintain the possibility of ordinary predication about ourselves. Am I myself then? If I think that I am the subject of my conscious experience and the cause of everything I do, it seems doubtful that Dennett's narrative view can answer affirmatively. The self, as he conceives it, is an illusion produced by the organism, and has no causal power. It is a functional property of the organism that allows it to simplify the organization of information. The self is not the subject of experience, it rather seems to be a feature of the experience that the organism has. In addition, such experience seems to be merely illusory. And as I noted earlier, a functional or psychological property is just not the sort of thing that can own my clothes or like ice cream. This sets Dennett's fictionalism very much apart from those metaphysical views I presented earlier.

3.1.2. The Narrative Self: from metaphysics to psychology

How does Velleman's conception of the narrative self relate to Dennett's?

[22] For a discussion of the problems with this distinction between encoding and exemplifying for our treatment of fictional characters see ibid.

As we have seen in chapter 2, Velleman claims that we should consider the self as a narrator and the narrative as an autobiography that feeds back into behaviour by means of actual agential control. Dennett's view that the self is an illusion strips it from having any causal role. If we were to reintroduce causal powers to a self conceived as purely fictional, we would be in an alarming position of considering a non-concrete entity as endowed with effectual and concrete causal powers.

However, Velleman's position does not actually lead to this result and it can, in fact, provide a narrative account devoid of the metaphysical ambiguities I highlighted in Dennett. Dennett's account seemed to fluctuate from considering the self as a fictional object to describing it as a function of the organism and, as we have seen, this produced some ambiguities in the way we can even talk of ourselves as selves, since the organisms' functional properties do not seem plausible candidates to be metaphysical subjects. In fact this confusion between properties and subjects is not present in Velleman. He straightforwardly identifies the self with the agent's reflective narrative representation which allows the kind of control over one's own actions that amounts to acting for reasons. In this view, the self is decisively something I *have* rather than *am*: it "denotes a self-conception, rather than an entity, real or imagined."[23] I am not the same as my self: I am the unified agent who has a self, in the form of an "inner locus of agential control", which is a representation that allows the process of unifying my motives by weighing them as reasons and allows me to decide among them. This means that there really is one subject of experience and ordinary predication. However, the agency this subject enjoys can amount to behaviour guided by reasons depending on the process controlling it. This view is neutral about the metaphysics of the self conceived as a subject of experience, or a bearer of properties such as personal identity. In this sense, rather than one's *self,* if we reserve this name for the metaphysical subject, Velleman's view seems more accurately described as one that accounts for one's *self-representation*, where this

[23] Velleman, 2006, p. 214.

is the reflexive representation of some aspects of oneself to one's own mind.[24] This understanding of the self as a self-representation follows the conception of the self assumed in much of the work done in social psychology. [25]

Other authors share this conception of the narrative self (for example, Goldie, to be discussed shortly). Nevertheless, the metaphysical commitments of those who hold narrative views vary. After all, in my narrative self-representation there is an 'I' that figures in my stories and there can be different ways in which we understand its metaphysical nature and the relation it has to this self-representation. In order to further specify Velleman's position I will briefly present two philosophers who focus on the value of narratives in our agency and in our experience. The first, Marya Schechtman, assumes that this self-representation entails certain metaphysical characteristics attributable to the agent and could potentially fix the conditions for metaphysical subjects. The second, Peter Goldie, is entirely neutral about the metaphysical nature of such subjects.

According to Marya Schechtman, narratives are fundamental for one's survival.[26] Schechtman seems committed to a psychological understanding of personal identity and thinks that narratives are essential for any subject to have a "stable self", conceived in terms of his personality and defining traits that ground the possibility of there being unity of consciousness. Through narratives one is able to have empathic access to one's past experience and therefore narratives serve the process of *identification* with one's stable self.

On the other hand, Peter Goldie suggests that, in fact, we should not even talk of a narrative self, but rather of a narrative sense of self.[27] A narrative sense of self is a way of thinking about oneself, about my past or my future, as already a person. My narrative sense of self presupposes that there is a person, I, who does the thinking, and my self narratives are not mine in any deeper sense (for

[24] Questions about one's sense of self arise from phenomenology rather than from metaphysics. An attempt to ground a metaphysical claim on the phenomenology of the self can be found in Strawson, 2002.
[25] Velleman, 2006, p.13.
[26] Schechtman, 1996.
[27] Goldie 2012, pp. 117-118.

example, they are not, like in Schechtman, the conditions for experiences to be actually my *own*). Goldie is entirely neutral about the metaphysics of persons, but he also thinks that our capacity for narrative thinking is distinctively human and that a narrative sense of self is fundamental for our self-reflective abilities. Goldie in particular stresses that identification is not necessary for our narrative sense of self, because I can recognize as part of my story experiences from which I now feel completely alienated. I can have, towards them, different attitudes: for example, I can feel regret or look at them with dramatic irony. These attitudes matter to us precisely because there is no threat to the actual survival of their subject and our talk of someone as "not the same person anymore" in virtue of some change of his defining traits and stable dispositions must, Goldie argues, be understood hyperbolically.

Velleman's view seems to be an intermediate position between these two. On the one hand he assigns no special role to one's narrative understanding for issues regarding personal identity. In fact, his philosophical project over all can be considered as an effort to distinguish the different "guises"[28] in which we talk about the self.[29] The narrative module, which functions as the agent's central controller, provides a unity that is not to be understood as the unity of consciousness related to personal identity: "that which makes us self-governed is not that which makes us self-same through time".[30] It is in this sense then that we can understand how Velleman's appeal to narratives is very different from Dennett's, who was interested, for example, in establishing the conditions for the survival of the self and equated them with those we can attribute to fictional characters. Even if we were to doubt the extent to which Dennett's can be considered a view of personal identity, in Velleman's picture there are no doubts: the fact that someone *has* multiple self-conceptions controlling her behaviour has no implications for personal identity. In this sense, then, Velleman's position is similar to Goldie in its being neutral in respect of the metaphysics of the subjects. On the other hand though, Velleman ascribes a more ambitious role to narratives than Goldie: according to Velleman, our

[28] Velleman, 2006, p. 1.
[29] See 2007b and ibid., pp. 170-202.
[30] Ibid., p. 22.

narrative self-representation is fundamental for the very possibility of our full blown agency. The role of the self, conceived as providing agential unity, grounds the possibility of self-governance, self-control and what can be considered as autonomous agency.

These same considerations can apply to Korsgaard's conception of the self, which in her case coincides with one's identity. Korsgaard claims that one's identity is an activity and not a state: this makes our self-constitution as agents into a task constitutive of our nature. In her view, for example, being a giraffe, or being a cat, are activities, and as such they presuppose some entity that does x, y, z (like search for food, escape predators etc) in order to keep being a giraffe. Similarly, for us, being a self is an activity sensitive to constitutive standards. Any activity requires a subject, and in this sense it does not make sense to talk of the property of being a self if it is not attributed to a subject. The very talk of identification, which as we have seen is a feature of experience, presupposes some subject enjoying such an experience. The talk of selves, then, assumes that there is an entity the self-representation is a representation of, or, in Korsgaard's case, whose activity (of identification) it is.

To sum up then, Velleman's theory conceives of the narrative self as a particular self-representation and can be in principle metaphysically neutral in respect of the nature of the subject enjoying it. This seems to remove some of the ambiguities that Dennett's original definition of the narrative self presented, because it clearly makes the self an element of one's psychology and does not link it to metaphysical claims. Nevertheless, Dennett's view, from which Velleman takes the lead, also considers that the self is ultimately an illusion conjured up by the organism. I will now discuss in what way this claim can result in a challenge to Velleman's view and whether he can provide a convincing answer.

3.2. Is the Narrative Self an Illusion?

So far, we have seen how Velleman takes the lead from Dennett in thinking that selves are better understood as fictional, but he also makes the explicit claim that the self, as a self-representation, is an element in the persons's psychology. According to Velleman, then, this narrative module is functional for our cognitive economy and it enables our behaviour to be self-controlled. Because of this, Velleman opposes Dennett's conclusion that, since one's self-representation has been conjured up by the organism, this automatically makes one's idea of being a self-controlled agent false. In Dennett, the fictional character of the self made its causal power merely illusory: in fact, the metaphor of narration enables him to suggest this idea, because narratives are not just reports of facts. Dennett's view assumed that, since one's self-narrative is fictive, it is false. The position Velleman wants to advocate for is seemingly paradoxical: the narrative which grounds agential control is in fact both fictive and true.[31] This means that, for Velleman, one's fictive self-narrative is factual and as such it is not a mere illusion. It is fictive, because it is made up by the person, and it is factual because it is in fact what the person actually enacts. Velleman concedes that "to be sure, a self-narrator can go beyond what is factual, if he applies self-descriptions whose autobiographical application will not make them true".[32] In these cases the self-representation is not effectual in determining one's behaviour. However, the fact that these failures are possible does not mean that the self and its role in determining our behaviour are merely illusory. Such a claim would be unwarranted. In fact, considering these cases as ones of *failure* entails that the normal functioning of these mechanisms allows the production of behaviour with the distinctive characteristic of being self-governed, since it is regulated by one's own self-representation. So, contra Dennett, Velleman holds that it is possible for the kind of creature that we are to produce behaviour according to our self-representation that functions as 'central meaner and controller'. However, Velleman has not yet positively established that this is in fact the case and in what follows I mean to test whether or not he succeeds in doing so.

[31] Ibid., p. 203.
[32] Ibid., p. 21.

As we have seen, Dennett tried to question the folk-psychological intuitions about mental causation, which conceived the self in various ways (for example as a soul, a ghost in the machine or some part of the brain) but which saw it as a "benevolent Dictator ruling from Headquarters".[33] Velleman, by contrast, aims at making sense of such intuitions so that he could explain how not only our beliefs and desires are causes of behaviour, but that we, as agents, are: this is because we can, through our narrative module, voluntarily select and endorse which mental properties are going to guide our behaviour and control the process. Moreover, the model considers that the process can be consciously tracked and couples it with our experience of agency. Velleman, as we have seen in the previous chapter, seems to give two conditions for one's narrative self-representation to be factual and thus productive of behaviour: on the one hand, actions need to result from the right psychological structure (one regulated by one's narrative self-representation); on the other hand, cases of INS*, in which one's representation in fact controls behaviour but is not accessible, showed how one's conscious access is also a necessary component of actual agency. This means that in order to positively affirm that the self is in fact "both fictive and factual" and deny the conclusion that the self is an illusion, Velleman needs to show that these two conditions actually hold. I consider two challenges to this claim: the first one comes from much discussed evidence some theorists appeal to in order to make a case against the possibility of a 'conscious will'. I will show that Velleman can avoid this challenge. The second one questions the possibility that narratives can play any real 'factual' role, as Velleman supposes.

3.2.1 A first Challenge

As we have seen, Dennett challenges the very experience of agency and, in considering it as illusory, he deems these experiences to be systematically flawed and our judgements in the attribution of agency condemned to error. In fact, empirical work has been interpreted as challenging the possibility of

[33] Dennett, 1991, p. 416.

agency along similar lines as Dennett's. This is the line of thought followed by Daniel Wegner's studies that deny conscious mental causation.[34] According to Wegner, what he calls the "conscious will" is experienced when we infer, correctly or not, that our thoughts caused a certain action. He thinks that we draw this inference when we have a) thoughts that occur just before the actions, b) when these thoughts are consistent with the actions, and c) when other potential causes of the actions are not present. Like Dennett, Wegner thinks that our actions actually spring from sub-personal processes and that those thoughts we experience as causes of such actions are themselves products of processes at the sub-personal level, which may only have indirect link to the causes of actions. Wegner concludes that when agents interpret their experience of agency, all ascriptions of agency are the result of their confabulation and rationalization. The "conscious will" plays no actual causal role. In the light of these considerations, one could challenge Velleman's picture in the following way: Velleman claims that our narrative self-representation results from cognitive drives of self-understanding. These drives are in fact sub-personal processes that do not guarantee that the conscious self-representation they originate is actually connected to the processes that produce behaviour. Therefore these conclusions deny a connection between our experience of agency and the psychological structure that underlies action. But this is precisely what Velleman requires to set the conditions for genuine autonomous agency. As a result, the claim that the experience of agency is entirely illusory bears on the possibility of action altogether.

There is room for Velleman to resist this challenge because the kind of agency required by the narrative model for action does not, contrary to appearances, correspond to Wegner's experience of a "conscious will". Since Velleman's conception of agency does not require the kind of experience that Wegner's project aims at debunking, Wegner's work does not present a challenge to Velleman's picture. I will briefly say how this is the case.

[34] Wegner, 2002.

Wegner thinks that, for our idea of a conscious will to be true, one would need to be endowed with a metaphysically free will, understood as an uncaused cause of action.[35] He presents the results of his empirical work as evidence for determinism and believes that this kind of metaphysical freedom would be incompatible with it. However, Velleman is not committed to the view that agency necessarily requires metaphysical freedom. Instead, he differentiates between this and epistemic freedom:

> On the one hand, there may be no particular way that the future is going to turn out— or at least, no way that's necessitated, under the laws of nature, by the present state of the world. In that case, the future would be causally or metaphysically open. On the other hand, there may be no particular way that we must describe the future as turning out, in order to describe it correctly—or at least, no way that's necessitated, under the laws of nature, by a correct description of the present state of the world. In that case, the future would be, as I put it, epistemically open.[36]

I will not analyse Velleman's distinction between epistemic and metaphysical freedom in detail[37] but I will try to show how we can understand Velleman's claim that the experience necessary for actual agency is connected to epistemic rather than metaphysical freedom and is not, therefore, incompatible with determinism.

To understand what it means for us to be epistemically free Velleman considers the example of someone ordering lunch at a restaurant.[38] In this situation there are many answers to the waiter's question about what you will have for lunch, and if you say any of those alternatives, you will indeed have it

[35] Several commentators have noted (Bayne, 2006; Carruthers, 2007) that Wegner seems to have a number of different illusions in mind. The experience of freedom and the experience of conscious mental causation appear to be his main targets.
[36] Velleman, 2000, p. 34.
[37] This would require me to discuss the notion of epistemic freedom in more depth than it is necessary for this discussion. Such an enterprise would need an analysis of connected claims, such as the purported difference between commands and self-fulfilling predictions (Ibid., p. 49), the distance of this model from one that attributes the impression of openness of future to a state of ignorance (Ibid., pp. 38, 50) and more generally Velleman's conception of intentions as self-fulfilling beliefs, which has been much debated and reconsidered by Velleman in different places (see Velleman, 2007a, p. xix; ibid., pp. 99-122 and 244-281) for a discussion see Holton, 2009.
[38] See Velleman, 2000, p.50. Velleman discusses two examples discussed, one is Anscombe's and one is about ordering lunch. I shall consider the latter only, so as to avoid discussions of the way Velleman interprets Anscombe's case.

for lunch. Answering the waiter's question makes you feel that your future is open. You feel your future to be open, in respect to what you'll have for lunch, because you know that there isn't one predetermined thing that you must say you are going to have: saying one thing or the other to the waiter in this situation will determine what your lunch is going to be. You will be correct about what you will have for lunch because what you are going to have is up to you: to your request, the waiter is not going to reply "Sorry, Madame, that is not actually what you are having" as if your lunch choice was not up to you. The alternative possibilities that we feel we are correctly entitled to affirm would be consistent with what he calls "central evidence" which concerns the reliable connection between our intentions and the outcomes fulfilling it. So, in ordering the sandwich, you assume that your assertion to the waiter will result in the ordered item to be presented to you. We have seen how actions were, in Velleman's model, the fulfilment of one's self-conception. In this sense they get performed only because the agent thinks he will perform them and wouldn't be performed if the agent thought otherwise: in light of this, Velleman claims that our actions are epistemically free.
Epistemic freedom does not, however, imply that you are not physically or psychologically determined: it is in fact not causally possible, according to Velleman, for one to do otherwise and we are predetermined to intend some thing and not another.

Thus, for Velleman, in acting we are in a particular conjunction of mental states, which include our knowing that we will bring about what we intend, within our possibilities, and our desiring that some course of events is brought about. In these cases our behaviour is regulated by those cognitive drives of self-understanding and self-consistency that defined agency: these drives bring about the prediction our self-narrator makes, and our behaviour still counts as an autonomous action. The result of this is that the process of producing action, as autonomous behaviour regulated by our narrative self-conception, can be understood as working mechanistically, in a way that is compatible with a deterministic framework. Even if determinism is true, agency would not be illusory. All Velleman needs for behaviour to be guided by one's

representation, and hence count as action, is the right psychological structure and conscious access.

To sum up, since Wegner's work attacks the possibility of an incompatibilist conception of free will, it poses no particular challenge to Velleman's picture of agency, which fits within the deterministic framework. This means that, in Velleman's model, the mechanism producing actions can be deterministic in the way the various psychological elements are connected; nevertheless, this still allows the agent to experience himself as autonomous and make him entitled to make decisions as self-fulfilling predictions. This allows Velleman to preserve the definition of action as behaviour guided by one's self-representation, in face of objections based on Wegner's work.

3.2.2 A second challenge

There is another way in which we can think that Dennett's claim about the narrative self being illusory might be problematic for Velleman. For the narrative self to be factual, as Velleman claims, it needs to play an actual role in action. Someone who claims that our narrative self is a mere illusion can then claim that the self never actually plays the role Velleman thinks, but we just have the illusion that it does. Remember Velleman's reasons why full-blown agency is not compatible with phenomena like confabulation and self-deception. The actions performed by confabulating and self-deceiving agents do not present the right psychological structure because one's behaviour actually does not result from one's cognitive drives reinforcing the motives that lead the agent to act. What is missing in cases of self-deception for them to count as cases of agency is a match between the agent's self-understanding and the actual motives driving the agent's behaviour. This means that although agency does require the agent's access and experience of the guiding self-representation, one's own conscious experience of such a representation is not sufficient for agency. In fact, in establishing whether some instance of behaviour is an action or not, we might end up contradicting the agent's experience. Our experience of agency is not reliable and we can consider our behaviour as self-governed and controlled, when in fact it is not. Attention to

these cases shows that our epistemic freedom, and the sense of agency that emerges from it, does not track only actions, and can be enjoyed in behaviour that does not exhibit the kind of psychological structure required for actions.[39] How do we rule out the sceptical worry that we could be always confabulating, or always self-deceiving and there are in fact no instances of full-blown agency? There is no real difference between the epistemic freedom and the experience of agency enjoyed in these cases and in full-blown actions.[40] Without establishing that the mechanism of cognitive drives postulated by Velleman is actually involved in action production, we cannot settle the issue of whether our narrative self, conceived as a narrative self-representation actually guiding behaviour, is illusory. If our self-representation was revealed to always be illusory, it could never match the actual motives driving behaviour. As a result, one's representation would in fact never play an actual role in guiding behaviour, and the mechanisms that Velleman thinks are necessary for agency would never actually be effectively at work. This would rule out the possibility that we are ever really acting.

Since this objection seems to be an empirical request for evidence of the phenomenon, I will next consider the array of studies Velleman appeals to support his view. These studies are not conclusive to support Velleman's narrative model because they can only provide some evidence to the idea of the function of the cognitive drives of self-understanding and consistency. I then turn to discuss more recent studies that have been developed to specifically support narrative views of agentive awareness in cognitive psychology. I will conclude that these empirical studies do not answer the question whether the narrative self is an illusion. This work has been conducted within the study of

[39] This need not be a problem for Velleman's claims about the instrumental value of this "useful fiction", but it is important to stress how it bears on the aims of Constitutivism. In fact, Velleman claims that his model can be the starting point for important conceptual work done on the ideas of responsibility and blame. An alternative proposal would consider that the same range of behaviour towards which we feel free is the one that we can be responsible for. This means that some behaviour that is not actually action will still fall into our responsibility, which would be more tightly linked to answerability than to causality. This alternative would clash with Velleman's causal account and his Constitutivism (inasmuch as it would make it unnecessary to track actions and give the criteria for different kinds of agency).

[40] What the model seems to fail to offer us is an explanation of the relation between the experience of agency and the psychological structure necessary for action. I will expand on this point in the next chapter, where I will consider the shortcomings of the definition of control and guidance in place in both Velleman's and Korsgaard's models.

agentive awareness and cannot provide an answer the question about the mechanisms of action production in Velleman's narrative model. I believe that the reason for this lies in the characteristics of the self-representation implied in Velleman's narrative account.

Velleman considers some classical studies[41] that confirmed Prescott Lecky's intuition that a person's conception of his world and of himself can play a functional role by organizing thought and behaviour into a unified whole.[42] The appeal to self-consistency would explain how one would perform (or refrain) from certain actions depending on how they can be assimilated into one's self-conception. Among others, Velleman considers phenomena like "cognitive dissonance",[43] whose evidence he interprets as showing how subjects conjure a motivationally relevant attitude to maintain explanatory and predictive coherence, and "self-verification"[44] that shows that subjects tend to retain feedback and behave in ways that confirm their self-conception. Studies comparing attribution and persuasion as means of modifying the behaviour of children showed a significant difference between those groups subject to persuasion (who were told that they *ought to be* tidy) and groups that were subject to attribution (who were told that they *were* tidy). The considerations drawn from experiments of this kind pushes in the direction of seeing subjects behaving in ways that verify self-attributions.[45] Velleman appeals to this to confirm his idea that cognitive drives of intelligibility and consistency play a

[41] Among others Nisbett and Valins, 1972; Nisbett and Wilson, 1977; Bem,1972.

[42] Lecky (1945) built his hypothesis around examples of students whose failure in certain subjects (spelling) was, he thought, due to their active resistance to overcome their difficulties. This resistance would come from the fact that they have assimilated the idea of being poor spellers into their definition of themselves.

[43] Velleman considers different experiments, from the classic studies conducted by Festinger and Carlsmith (1959) to Aronson (1969).
While not providing positive evidence for the actual causal role credited to this cognitive drive, Velleman considers this evidence to support what he calls "a complicated reality" (2006, p. 233). Next to the possibility for interpretation to fit to behaviour, there is the chance of behaviour to fit to interpretation: this research can be taken to show how episodes of confabulation and rationalization in these experiments confirm the idea of a cognitive drive for self-understanding that, in these cases, catches up with behaviour that was performed in absence of adequate understanding.

[44] Swann,1986; Miller, Brickman and Bolen, 1975.

[45] Velleman discusses various variations on the present experiment see 2006, pp. 236-237.
Velleman stresses how these experiments cannot be conclusive in supporting Lecky's (and his) hypothesis inasmuch as the attributions studied are all positive and could hence be explained by a phenomenon of self-enhancement. Moreover, he points at how there has been intermittent success in duplicating results involving a negative self-conception.

crucial role in action: these drives motivate the process of self-interpretation and guide action by fitting behaviour to one's self-understanding. We have seen already how this means that, according to Velleman, these drives "play the role of the agent".[46] Nevertheless, these studies do not directly support the idea of the role played by narrativity in agency and therefore cannot provide an answer to the question about the illusory role of the narrative self.

In more recent work, the importance of narrativity has been stressed to explain agentive awareness, by theories that seem to retain many of the elements of Velleman's account. Theories in psychology that ascribe a central role to a narrative module in agency all focus on high-level integrative processes which result from a central, domain-generic, holistic mechanism.[47] These narrative models are top-down approaches that emphasize the role in agency of high-level, comprehensive and conceptually-laden intentions and beliefs as opposed to low-level sensorimotor cues, for example in agency ascriptions and agentive awareness. I will then now turn to these studies in order to establish whether they can support Velleman's view.

Neurological work in support of this approach can be found in Gazzaniga's[48] experiments on split-brain patients: he showed that subjects are prone to confabulate accounts of actions that are generated only by the right hemisphere. This led Gazzaniga to claim that the left hemisphere can be credited with a mechanism which he calls 'the interpreter', responsible for generating narratives, and compensating for missing information. In psychiatry, prominent narrative interpretations on schizophrenia[49] consider delusions of alien control to result from the fact that these thoughts and actions may not make sense in relation to the whole, thus conceiving of agentive awareness as depending on the subject's "ability to integrate these thoughts and actions into her larger picture of herself".[50] The process of interpretation, so described, seems to amount to "turning our mind-reading capacities upon

[46] Velleman, 2000, p. 137.
[47] Bayne and Pacherie, 2007.
[48] Gazzaniga and Ledoux, 1978.
[49] Sass, 1992; and Stephens and Graham, 2000, elaborating Sass.
[50] Stephens and Graham 2000, p.89.

themselves"[51]. In general narrative approaches hold that, while the subject's judgement in attributions of intentions and agency depend on his experiences of agency, they place substantive constraints on whether or not the contents of these experiences are to be accepted. In this sense, they can exercise the ability of veto and thus negate agency for experiences that the subjects cannot integrate into their self-conceptions. In a similar fashion, we can also assert that we are the authors of events for which we lack such an experience.[52]

It is difficult to find support in this research for Velleman's view that people are "generally guided in behaviour by a cognitive motive towards self understanding"[53] since these studies track the phenomenon of agentive self-awareness and are concerned with the conditions of the emergence of the agent's sense of agency. Even granting that the narrative approach is right about how the sense of agency emerges,[54] the connection between this phenomenon and how things are caused is still unclear. In fact, mechanisms responsible for the generation of the sense of agency are only indirectly related to the mechanism responsible for action production, and this results in dissociations between the sense of agency and agency, which oblige psychologist reconsider the role of the sense of agency in action causation.[55]

I believe that there is a fundamental reason why it is not possible for narrative models to reach a definite answer on whether the narrative module actually plays a causal role in action, namely, the nature of the self-representation the self is here conceived to be. I will discuss this in the following section. I think that Velleman does not fully consider the consequences of integrating a

[51] Carruthers, 2007, p. 199.
[52] Pacherie discusses the experiments conducted by Schultz *et al.*, 1980: a knee-jerk reflex was induced in children. "When asked whether they had meant to move their leg, three-year-olds said that they had meant to move their leg but five-year-olds denied it correctly…Top- down influences may therefore enhance otherwise elusive experiences through a process of attentional amplification. Thus, in younger children the constituent elements of the experience of control may already be present and yet go unnoticed, while older children with a better grasp of the concept of intention may attend to their experience, thus increasing its salience"(2007a, pp.21-22). Pacherie discussion is longer and critical in respect to top-down approaches, and in presenting it here I do not mean to commit her to a narrative approach.
[53] Velleman, 2006, p. 8.
[54] And there are significant doubts about this possibility, see Pacherie, 2010; Bayne and Pacherie, 2007.
[55] Pacherie, 2010.

narrative understanding of the self in his work on practical reasoning and agency. I will argue that because of the characteristics of this narrative self-representation, it is doubtful that Velleman's model can succeed in proving that the narrative self is not illusory and plays a factual role in action production.

3.2.3. The Characteristics of Narrative Self-representation

In this section I will specify what I think are the characteristics of a narrative self-representation. I believe that this characterization is entailed by all narrative approaches and bears on the possibility of establishing the illusory character of the self in Velleman's view.

Firstly, one's narrative self-representation is in general to be understood as a kind of judgement, a doxastic attitude expressing a belief, whose content is conceptual and which can include an array of different kinds of self-ascriptions. For example, one's narrative self-representation allows one to self-ascribe many different kinds of properties or mental states and events: one's conception of oneself under some social, cultural identity but also dispositional properties, psychological attitudes, empirical beliefs about oneself, as well as memories or plans about the future, can all figure in one's narrative self-representation. This means that one can represent oneself as loving his or her children, as happy, angry, guilty, as someone's friend or as a faithful husband, as well as believing that some position is wrong or wanting more cake or remembering having been to Vienna.

I take a narrative self-representation to be a judgement of conceiving oneself under some predicate, which results in a belief that is truth assessable.[56] So the constituent propositions of a narrative are true if they correspond to how things are. Moreover, narrative self-representation has explanatory significance, so that one's narrative self-representation explains my intention

[56] This is particularly important for Velleman, as we have seen in Chapter 2, cf. 2000, p. 25.

or action. My being someone's friend explains my concern about her current troubles. This means that we can attain "integrative knowledge of what we are doing simply by framing and fulfilling integrative conceptions of our behaviour, conceptions formulated in terms of the dispositions and circumstances that help to explain it".[57] On the one hand, this reference to an explanatory significance assumes that one's narrative self-representation requires self-awareness of being in a certain state. This means that, although not occurring, such narrative must be either conscious or capable of being accessed and becoming conscious, or else it would fail to count as a reason-giving state at all. On the other hand, in order for this explanatory relation to hold and frame these integrative conceptions at all, we need to be aware of the relevant factor with which we integrate our behaviour, the desires that might motivate it, the customs and policies it might implement, and in general the beliefs, emotions and traits of character it would express.[58] In narrative accounts these ascriptions have a certain meaning within a certain narrative: it is the narrative interpretation that gives them their significance.

The theory thus assumes that we ordinarily try to identify our behaviour at a "high" and "comprehensive" level, according to our "human inclination to be informed of what we are doing in the most integrative and general way available":[59] a knowledge of what we are doing in terms of why we are doing it. This means that my walking out of the pub is expressive of my being uncomfortable with my colleagues, of my intention of not wanting to have a drink with my colleague who's having an ambiguous attitude, of my caring about my husband and my being a faithful wife. In turn these descriptions all explain my behaviour. When asked why I am leaving I can give any of these

[57] Velleman, 2006, p. 264.
[58] Borrowing from Kant in another context, Peter Goldie has described this relation between the narrative self-representation and the pattern of expression in the following way: "a trait is the ratio essence of its pattern of expression; and the pattern of expression is the ratio cognoscenti of the trait" (2012, p.136). While I believe this is a natural way to understand the relationship, I also think that it would not really capture the conception of narrative self-representation now under examination. Firstly, and most importantly, because it does not account for the idea that the way in which one's self-representation can guide the production of an action is through the understanding of it that it provides. Secondly, the picture that Goldie provides seems to hint at an observational epistemic link which, as we shall see briefly, we need not assume.
[59] Vallacher and Wegner, 1985, p. 26.

explanations in a progression that displays high-level and comprehensive act-descriptions, to the point that I can answer the question about *what* I am doing directly with "I am a faithful wife".

Secondly, a narrative self-representation is not basic. It is an interpretative achievement, based upon other states and grounded in other more basic judgements or other subjective states. As already noticed, there are a wide variety of states that one's narrative self-representation can be based on. What is distinctive of narrative self-representation is its interpretive nature, which unifies all the relevant self-ascriptions in a unified interpretive structure.[60] It is not just awareness of being in some state. All the relevant states are organized in a structure which assigns them a certain meaning and bearing on the situation at hand: one's narrative self-representation interprets experiences and actions situating them in the context of a unifying story.[61] In light of what we have see so far, this means that one's narrative self-conception is a belief whose content is truth-assessable and is based on states which are themselves truth-assessable.

Moreover, the characteristics of the interpretive process are complex, as a result of the variety of different self-ascriptions that can figure in one's narrative self-representation. Here is an example, from Jane Austen's *Emma*, discussed by Crispin Wright, of how these attitudes can figure as part of such a process:

> Emma's eyes were instantly withdrawn; and she sat silently meditating, in a fixed attitude, for a few minutes. A few minutes were sufficient for making her acquainted with her own heart. A mind like hers, once opening to suspicion, made rapid progress; she touched, she admitted, she acknowledged the whole truth. Why was it so much worse that Harriet should be in love with Mr. Knightley than with Frank Churchill? Why was the evil so dreadfully increased by Harriet's having some hope of a return? It darted through her with the speed of an arrow that Mr. Knightley must marry no one but herself![62]

[60] See Zahavi, 2007, p. 179.
[61] This is the case regardless whether such a story is a life-long unity or a short episode.
[62] Austen, 2009, p. 249 quoted in Wright, 1998, p. 17.

In this passage we have a good example of a fairly common process of self-interpretation that Wright contrasts with our common knowledge of states like 'I am tired' or 'I am hungry'.[63] Without any special commitment to ascribing some special knowledge to these states,[64] we generally consider one's judgement of being hungry as a reasonable self-ascription by the subject having an experience of hunger. Similarly, we would allow Emma a similar form of authority in relation to her judgement of being angry based on the experience of being angry. Her conclusion about being in love is drawn from considerations of this sort which are steps in her interpretive process for which we can see her retaining special authority. So Emma considers that she is angry at Harriet, or that the thought of Mr Knightley returning her friend's love is painful, or that she thinks it is much worse that Harriet is in love with Mr Knightley, than with Mr Churchill, etc. She thus recognizes that she must be in love with Mr Knightley herself.[65] While we consider that her being in love is a reasonable explanation of her other attitudes, we see not only that an external observer might have drawn the same conclusion, interpreting some different evidence (for example, based on Emma's behaviour), but that clearly

[63] Wright calls such self-ascriptions 'phenomenal avowals', of which, he considers, we have a knowledge that is groundless, inasmuch as any demand for reasons or corroborating evidence - any question such as "How can you tell? - would be inappropriate; authoritative, inasmuch as the subject's taking that he exemplifies a certain mental attribute is a guaranteed indication that he does; and transparent, in that if you consider that you are scared, tired or hungry, it is expected that you will know it. See Wright, 1998, p. 16.

[64] In this I wish not to endorse, for example, any introspective view (Goldman, 1993) or "no-reason" view (*a là* Shoemaker, as discussed in Peacocke, 1998).

[65] One might object that Emma's example does not capture the way in which we normally make such self-attributions. For example, Velleman has already stressed that our self-attributions remain unarticulated (2007b, p. 281) and is not committed to consider them as appearing in our occurring thoughts. Surely, for example, we do not normally follow Emma's procedure in respect of our intentions. This reply would misunderstand what emerged from the discussion of the example. The non-basic character of one's self-representation and the way it is based on other subjective states is a matter of the representation's nature and structure rather than a claim about the way (implicit or explicit) in which one's narrative self-conception is represented. In fact, in Velleman's evidentialist epistemology, one's narrative self-representation remains in the background of thought, a tacit knowledge on which the knowledge of our own actions and intentions rests on without being 'derived from it'. Velleman considers that the knowledge of our actions rests on the background knowledge of our surroundings and abilities, so: "In order to know that you are opening the window, you have to know that it's a window and you can open it" (2007a, p.20). In knowing what you are doing you rely on some background knowledge and your narrative self-conception is one of its pieces: it specifies the description of your actions in the first place. As I explain in what follows, I do not see the need of committing to any specific way in which we can acquire knowledge of our narrative self-conception: whether explicitly or implicitly, whether inferentially, observationally or immediately. For a discussion of Velleman's inferentialism see Roessler, 2013.

she could be totally mistaken (she might, for example, just be lonely and possessive in respect of this long-term friend). Similarly one can consider that my self-representation as a "faithful wife" is based on my interpreting my intention to leave the pub, my desire not to have another drink with my colleague and my feeling discomfort in the situation. One's self-interpretation is based on self-attributions of attitudes which can themselves be subject to misunderstanding or insincerity, as a result of self-deception or wishful-thinking. So, my consideration that "I don't want another drink" allows for misinterpretation in forming this attribution: I might in fact be attracted to my colleague and be mistaken about my higher-order attitude, interpreting my emotional turmoil as discomfort rather than attraction. In this sense my self-representation is the result of an interpretive process that is largely open to mistakes and is grounded on more basic elements that themselves could be mistaken. In pointing at these epistemic difficulties we need not to assume that one's self-interpretation needs always to be conducted at the personal level or needs to be explicit, based on introspection or inference, as in Emma's case.

In fact we need not consider this self-interpreting process as inferential at all, although it is important to notice that most advocates of narrative approaches seem to consider it as such and some explicitly say that "we infer the existence of propositional attitudes in ourselves in much the same way that we infer their existence in other people".[66] Indeed, some[67] seem to invoke the role of self-

[66] Stephens and Graham, 2000, p. 161.

There is one way of understanding the considerations I drew from Wright's discussion that could lead to some misunderstanding. Wright stressed how Emma's conclusion resulting from an interpretive process could be seen as knowledge that did not preserve some fundamental asymmetry between self and other. Indeed, one's narrative self-representation seems less private and more complicated than knowing the immediate beliefs and desires on which it is based.

This means that "our self-making stories are not made up from scratch; they pattern themselves in conventional genres…When I interpret my life story, I might be both the narrator and the main character but I am not the sole author" (Zahavi, 2007, p.181). In particular, we can consider how the asymmetry between one's own narrative self-representation and others' is contingent on the extent of knowledge: we just happen to know more about ourselves, and so my narrative is more replete, but this needs not to necessarily be the case at all. "In principle the content of my narrative self-representation can be isomorphic with the content of your narrative sense of me" (Goldie, 2012, p. 121). What is expressed by me when I self-ascribe "I am a loyal friend" is by no means private or inexpressible by others and can in fact be formulated in the second or third person. Moreover, others play a substantial role in accepting the narrative accounts we give of our actions and experiences. This is why we ask counsel and consider that others might have revealing insight in

interpretation, as a process relying on evidence and inference, even for some basic self-attributions like "I am angry/scared/happy".[68] In light of these considerations, Emma's case can be seen as paradigmatic of what is normally conducted sub personally. In particular, if we consider that the building blocks of experience are fragile and recessive, they might postulate that no trace is left of the very basic evidence on which self-attributions are based. However, I do not think it is necessary to commit to any substantial position about the introspective or inferential character of the knowledge of our narrative self-representation, in order to claim that the high-level, integrative self-representation that results from this interpretive feat allows the possibility of severe misinterpretation.

In fact, the possibility of misinterpretation is not much linked to a supposed inferential character of one's narrative self-representation; it rather emerges from the vast variety of self-ascriptions that can figure in our narrative self-representation and whose knowledge conditions can be very different and elicit different intuitions. Some attributions of mental states and events might seem opposed to observational knowledge, others less so. The very placing of this self-representation at a conceptual and high level seem to posit substantial constraints in this sense: it employs concepts that are the result of social understanding as well my own personal interpretation of the roles they seem to ascribe. Arguably, when it comes to our dispositional attitudes or our social and cultural identities, we come to think of ourselves as friendly, honest and so on by taking a third-person perspective on our behaviour. Other self-ascriptions can be drawn by experiential memory and require a purely first-personal perspective on a particular event. But narratives can be compiled by

---

interpreting our own attitudes. In itself, the fact that there is no real asymmetry between self and other kinds of knowledge has consequences for the question of whether or not this narrative self-conception is illusory. Exactly because one's self-representation constitutes a judgement that is truth-assessable, one person holding self-deceiving beliefs is under rational obligation to explain away contrary evidence or else revise his beliefs: others play a substantial role in this revision. One's narrative self-representation would then be subject of rational standards in similar ways as other beliefs one can have.

[67] Stephens and Graham 2000, p.89.

[68] Velleman himself cites these classic studies on misattribution of emotions (2006, p. 243), reporting experiments conducted by Schachter and Singer (repeated by Zillman, Johnson and Day). In these experiments the subject's self-attribution of anger was manipulated by producing physical signals that the subjects misinterpreted.

different sources, and they could in fact include episodes I have forgotten and on which I have an external perspective, relying on the testimony of others.

To sum up, one's self-representation is necessarily an interpretative achievement and is based on more basic self-attributions which it combines in an integrative structure. The reliability of such interpretation depends upon the reliability of both more basic material and of the interpretive process itself. Even if we considered one's ascription of a narrative self-conception as spontaneous, immediate, not relying on inference or introspection, this would not deny its non-basic character and the fact that errors can involve different aspects that our self-conception integrates. Because of this distinctive character of narratives, we cannot rule out the possibility that our self-representation serves a compensatory function and responds to what Goldie defined as our "fictionalizing tendencies":[69] our need for satisfying coherence might distort reality by imposing fictional configurations on the facts and actions it strives to make sense of.

It is important to stress that the unreliability of narratives is not a worry that can undermine the whole project of explaining agentive awareness through a 'Narrators' Approach'. But it is a genuine worry for a model like Velleman's that closely links the motives causing behaviour to the reasons we endorse through our narrative understanding. Moreover, we can now see how some of the difficulties I have pointed at in the previous chapter are connected with this conception of one's narrative self-representation. We have seen how Velleman's model ultimately restricts unduly the range of actions. This is the result of the implausible standards of self-understanding that Velleman imposes. On the one hand, the high level of sophistication that narrative self-representations imply is bound to constraint the range of actions the model accounts for. On the other hand, we can see how these standards of self-understanding are in tension with the reliability we can attribute to narrative self-representations.

[69] Goldie, 2012, p.161.

To sum up, because one's narrative self-representation is placed at such high level, there is constitutively room for misinterpretation. The possibility of self-deception and error cannot in principle be ruled out and this affects the way in which we can consider the relation between the actions that should be expressive of such a self-representation. Given these constitutive epistemic difficulties, it is possible to understand why top-down narrative approaches cannot provide evidence that can settle the question about the illusory character of the self.[70]

The kind of self-representation which, according to Velleman, guides one's behaviour in actions has been shown bound to produce the shortcomings I highlighted in the previous chapter: while Velleman's narrative model excludes that self-deceiving and confabulating agents are truly full agents, it is the very nature of narrativity that constitutively makes room for the fictionalizing tendencies grounding these phenomena. In what follows I want to examine a different aspect for which the conception of this self-representation is problematic. As we shall see, this discussion will lead me to question one of the fundamental theses that I attributed to both Korsgaard and Velleman: the idea that there are two levels of selfhood.

3.3. Two levels of selfhood?

I will now turn to consider a fundamental problem that emerges from the understanding of self-consciousness presented in these theories of reflective endorsement. I would like to step back and discuss how the understanding of this kind of self-representation can cast light on the commitments involved in what I called thesis (ii): the idea that there are two levels of selfhood. In this way we can connect the above considerations about narrative self-representation to Korsgaard's view. I presented both models as associating two kinds of agency with two kinds of subjectivity. This was possible because the distinction between kinds of agency ascribed a fundamental role to self-

[70] For an analysis of concerns of reliability associated with top-down and narrative approaches see Pacherie, 2010, pp. 453-454.

consciousness, considered as a distinctive human capacity. Both thus considered that the "self", as a distinctive product of the human capacity for self-consciousness, plays an essential role in defining the threshold of human agency, connecting *self*-consciousness and *self*-governance.

So, Velleman claims the following:

> As we have seen, self-awareness gives me an objective conception of the person who I am. That conception bears on practical reasoning, to begin with, by giving me access to objective knowledge of what I am doing.
> Of course, a cat is also aware of doing things, such as hissing at someone by whom it feels threatened. But a cat's awareness of its own doings never extends to the knowledge that they are being done by a creature in the world. It represents them from the perspective of the one doing them, without representing the creature occupying that perspective. Thus, even when cat is aware of hissing at you, and even if it is hissing with the thought of scaring you away, it cannot be thinking that you will be scared of this hissing creature - scared, that is, of its hissing self - because it has no conception of being one of the world's creatures, and hence no sense of self. By contrast, if I tried to scare you away, I would be aware of confronting you with a person saying "Scram!" as would be manifest in that very utterance, since a person saying "Scram!" is intimidating precisely by virtue of manifesting the intention to be an intimidating person.[71]

And we can read Korsgaard as agreeing with Velleman when she asserts:

> I believe that human beings differ from other animals in an important way. We are self-conscious in a particular way: we are conscious of the grounds on which we act, and therefore are in control of them…This means that although there is a sense in which what a non-human animal does is up to her, the sense in which what you do is up to you is deeper. When you deliberately decide what sorts of effects you will bring about in the world, you are also deliberately deciding what sort of cause you will be. And that means that you are deciding who you are.[72]
>
> The reflective structure of the mind is a source of "self- consciousness" because it forces us to have a *conception* of ourselves…It is better understood as a description under which you value yourself, a description under which you find your life to be worth living and your actions to be worth undertaking. So I will call this a conception of your practical identity. Practical identity is a complex matter and for the average person there will be a jumble of such conceptions.[73]
>
> Self-consciousness is the source of reason. When we become conscious of the workings of an incentive within us, the incentive is experienced not as

[71] Velleman, 2006, p. 258.
[72] Korsgaard, 2009, p. 19.
[73] Korsgaard, 1996, pp-100-101.

> a force or a necessity but as a proposal, something we need to make a decision about.[74]

Both philosophers allow that we can sometimes act in a way that is animal-like. When behaving in this sort of way, we are able to recognize that we are the cause of that behaviour: I know that it is me who knocked down the inkstand, and I know that it is me who shouted at my friend. Behaviour that does not count as action, though, is not *owned* at what Korsgaard called "a deeper level", the level that matters for human self-governance and that exhibits the distinctiveness of human self-consciousness. What I called the 'two levels of selfhood thesis' claims that one's objective self-conception is distinctive of human subjectivity and human agency. And this makes it a condition for this distinctive sense of ownership and authorship.

However, the formation of such self-representations seems to require the proper functioning of a variety of cognitive capacities, and it is to this aspect that I would like now to focus attention. This will lead me to identify the internal conditions for self-consciousness identified at the level of full action. I will start from Velleman's position, and then show how these considerations apply to Korsgaard as well.

Following from the previous discussion we can now see how, according to Velleman, one's narrative self-interpretation provides one of those third-personal ascriptions that are characteristic of this level of selfhood, as opposed to the egocentric perspective that allows animals to be conscious but not self-conscious. When one's behaviour is guided by such considerations and it is made sense of in light of them, one considers oneself reflexively as the subject guiding one's actions. In this sense one's objective personal conception and understanding enriches one's first-personal perspective. Now, the formation of self-narratives seems to require at least these two internal conditions: the

[74] Korsgaard, 2009, p. 119.

capacity for minimal self-awareness and the capacity for engaging in reflective metacognition.[75]

The capacity for minimal self-awareness is necessary because self-narratives are possible only if one is able to refer to oneself by using the first-person pronoun: this gives a secure point of reference for creating a self-narrative. Velleman holds that we are in general *unselfconscious* about the reference to oneself, when engaging in genuinely first-personal thought.[76] This means that one does not need to fix a target of self-reference: it is assumed. According to Velleman self-reference here is elliptical.[77] This secures the fact that I cannot be mistaken in my self-reference. Consider Emma's case again: we can see no attempt to identify oneself, because in her self-attribution the reference is fixed. In fact, as we have seen, Emma might be mistaken in attributing a set of properties to herself: she might be wrong about being in love. In this sense it would make sense to ask "Are you sure that you are in love?"; what would not make sense to ask is "Are you sure that it is *you* who is in love?". Although she can be mistaken about the self-attribution of a certain set of properties, she cannot be mistaken about the self-reference: indeed we can consider this a case of misattribution only because she has correctly self-referred.[78] In this sense, as

[75] Here I am following Gallagher, 2003. Gallagher discusses what he considers are the four internal conditions for self-narratives. Next to the two I discuss here, he cites the capacity for temporal integration of information and the capacity for encoding and retrieving memories episodically. I considered these aspects not immediately relevant for the present discussion, in light of the characteristics of Velleman's conception of narrative highlighted in Chapter 2.

[76] As opposed to when we are imagining being in someone else's shoes, for example, see Velleman, 2006, pp.170-202.

[77] Ibid. p.276.

[78] In its being linked to minimal self-reference, one's narrative self-conception is different from another form of self-understanding that Velleman describes: the way in which one recognizes oneself in the mirror. Velleman introduces this subject when trying to make sense of the specificity of human self-consciousness but it is unclear how we are to understand the relation between one's narrative self-conception and mirror recognition, which, Velleman says, requires the subject to have non-reflexive knowledge about himself. Velleman stresses that the recognition establishes a relation between reflexive and non-reflexive information about myself and it requires the idea of my self as a person, which is associated with my body but also with my reflexive methods of introspection and thought-control (2006, pp. 278-280). I think that one's self-ascription of a narrative self-conception (like Emma when she concludes that she is in love) is essentially different from the kind of self-representation necessary to pass the mirror test in the way that Velleman describes. In the case of identifying oneself in the mirror we can understand the following inference to hold.
"That person has x,y characteristic"
"That person is me"
"I have x.y characteristic".

a condition for autonomous agency, we see that there is a basic reflexivity that we need to associate to one's self-understanding, so that one's reflective objective personal perspective about oneself can be integrated with the essentially first-personal, and genuinely first-personal, standpoint of the agent.

The capacity of reflective metacognition enables the interpretational process that we have discussed already: narratives require reflective consideration of certain events and self-attributions. This means that one understands these events and ascriptions by fitting them together semantically: by assigning a certain significance and meaning to them. A narrative is more than a simple chain of remembered events or causal relations and one's objective self-representation shapes self-attributions into a narrative, enhancing the product delivered by self-ascriptions.

These considerations seem to apply to Korsgaard's view if we consider the characteristics necessary for the phenomenon of "identification" that is central in her model. Her conception of human self-consciousness also pulls together the essentially first-personal perspective of the agent with the third-personal self-conception of one's identity. This idea requires both the agent's minimal self-awareness and reflective metacognition: identification is not the mere ascription to oneself of some practical identity. It considers one's practical identity "as a description under which you value yourself, a description under

---

In this case I take a third personal perspective while appreciating that it is myself who is present in this third personal way. This definitely clashes with the case of Emma as I have just described it, where the subject's self-reference is fixed rather than a matter of inference. The difference can be understood if we consider an example in which Emma would in fact be in the position to misidentify the subject of this attribution: in reading a letter she considers that its author was in love. She might then realize that it is in fact herself who wrote the letter and thus conclude that when she wrote the letter she was in love. In this case, it would make perfect sense to ask "Are you sure that it is you who was in love?". She might after all be wrong about her being the author of the letter, with the result that her attribution of having been in love would also be mistaken. In this sense there seems to be a fundamental difference between one's narrative self-representation and the kind of attribution we can draw from mirror cases because the latter, but not the former, seem vulnerable to error misidentifications. I note this because Velleman presents both aspects but is not clear about the relationship he takes them to have to each other, if any. While I will not explore this issue any further in this work it can be considered another aspect of the notion of self-consciousness that Velleman does not clarify.

which you find your life to be worth living and your actions to be worth undertaking".[79]

Both authors consider it necessary to link the first and third personal perspectives for two reasons: firstly, they claim that only the high level of self-consciousness so described places behaviour at the level of reason. Secondly, they also claim that only this level defines the specificity of human self-consciousness as opposed to the basic form we can credit some other animals with. I want to question both these points: I will claim that this model of selfhood mischaracterizes the specificity of minimal human reflexivity and assimilates it with a generic "egocentric framework". Because they overlook this aspect, both Korsgaard and Velleman push the threshold of the specificity of human self-consciousness up to this extremely high level of complexity. I believe that this level of complexity is unnecessary both to establish the specificity of human self-consciousness and to place human behaviour within the domain of reason.

Consider the following quotation from Velleman:

> Throughout the paper I assume that "first-personal" thought is not necessarily personal, in that it need not involve the concept of a person. Creatures who lack the concept of a person can nevertheless manifest behavior that is to be explained by their having egocentric representations of their surroundings – representations whose content cannot be expressed without the help of first-person pronouns. We cannot explain the stalking behavior of a cat, for example, except in terms of perceptions expressible as "There's a mouse in front of me", "I'm close enough to pounce on it," and so on. Yet the attribution of such first-personal thoughts to the cat does not imply that it thinks of itself, or of anything else, as a person.[80]

In this passage, Velleman starts off by claiming that a first-personal perspective need not involve the concept of a person, and this seems right to me. Nevertheless, when he explains in what way it is possible to have a first-personal thought without the concept of a person, he assimilates this to the kind of perspective that creatures with non-conceptual capacities have. Velleman is here making a clarification of our use of first-person pronouns in

[79] Korsgaard, 2009, p.20.
[80] 2006, p.180n.

explaining animals' behaviour: these attributions of first-personal thoughts require us to employ first-pronouns, which are unavailable to the animal itself. Does this mean that, in the case of beings like us, who are endowed with the capacity of actual first-personal reference, first-personal thought implies the reference to oneself as the person occupying such a perspective? This seems to be what Velleman thinks: if 'I' thoughts are specifically human, and the difference between our awareness and a cat's awareness of his doings is that the cat's does not extend to the knowledge that they "are being done by a creature in the world",[81] it then seems that my first-personal awareness expressed in 'I' thoughts requires a representation of the creature occupying my perspective.

Now, I believe that this view is a misunderstanding of first-personal awareness. If 'I' thoughts are conceived thus, one's subjective reflexivity, which does not require the representation of the person occupying one's perspective, is just assimilated to the egocentric framework that we can associate with creatures having nonconceptual egocentric sensitivity. In assimilating 'I' thoughts with the concept of person we have a view of self-consciousness that acknowledges only a high level of complexity. This view ignores the possibility that a more basic first personal awareness is a pervasive characteristic of all our conscious states and is already distinctively human. As a result, the 'two levels of selfhood' thesis gives a picture that opposes a non-conceptual egocentric framework to a full-fledged objective self-conception, with a substantial gap between the two. A consequence of this view is that it is not clear how self-consciousness, as a real conceptual capacity, emerges in ontogeny or phylogeny.[82] I believe that if we give an account of this dimension of self-consciousness that these views ignore, we would not need the level of sophistication and cognitive complexity assumed by the self-representation presented by these models to explain the specificity of human agency.

Korsgaard and Velleman hold that this notion of self-consciousness is necessary because only the high level of self-consciousness so described places

[81] Ibid. p. 258.
[82] See Bermudez, 2000, for the claim that we need this kind of explanation if we want to make sense of self-consciousness as a genuine psychological capacity.

our behavior in the domain of reason. I do not think that this is the case: in fact, this view overlooks the role actually played by self-awareness and first-personal reference in rational activity.

Velleman and Korsgaard link acting for reason to reflective endorsement, which provides the test that establishes whether something can count as a reason. One's objective self-representation, either in the form of a narrative self-representation or as a practical identity, gives the substantive evaluative standards for this test. Only attitudes that pass this test can count as my own and this means that the standards for self-governance and autonomy coincide with those of reason. The sense of authorship and ownership required by both Velleman and Korsgaard is clearly not just the basic self-attribution that belongs to 'I' thoughts. Again, I believe that this picture is misguided because it does not concern itself with a full account of the different dimensions of self-consciousness. As I said above, the objective self-representation that both models consider necessary for self-consciousness requires minimal self-awareness. Both views remain elusive when it comes to define this minimal self-awareness, and focus on the distinction between an animal's non-conceptual egocentric framework and full-fledged human objective self-consciousness instead. I believe that they overlook the fact that the specificity of human reason can already be understood at a level of self-consciousness in which one's objective self-conception plays no role.

Thanks to a basic self-awareness and purely first personal thought, I recognize that some attitudes or actions are mine without the concept of the person that entertains those attitudes to play a role. This is not a kind of self-awareness that we can attribute to a cat but neither it is one that requires my objective self-representation to play a role. In first-personal reference one recognizes a thought as one's own and the basic ownership so expressed is the conceptual basis for any immediate rational relevance. When I self-attribute an attitude in the basic sense of having 'I' thoughts, I fix the locus of responsibility that is necessary for reasons. The first-person reference attaches a judgement to a person making the subject sensitive to the demands of rational evaluation. This is the case even when the subject responds to these demands by disowning his

actions in the deeper sense that Korsgaard and Velleman require. This step is prior to any full-fledged reasoning about one's attitudes through reflective endorsement. While one's objective self-representation can provide one with the evaluative standards to judge certain considerations as reasons, one's first personal awareness is all that is needed for one to be sensitive to the demands of reasons. In light of this, there is no need to assume that the basic conditions for reason-responsiveness are necessarily linked to a level of self-consciousness that requires the concept of a person and even less the objective self-conception of the person you are.

To sum up, I believe that placing the distinction between human self-consciousness and animal consciousness at the high level required for an objective self-representation does not provide an understanding of more minimal forms of self-awareness that are already specifically human and are involved in human agency. This conception of consciousness explains some of the shortcomings of the models highlighted above, because it bears on the account of reflection the models can give and on the range of actions they define.

For example, we have seen how Korsgaard's model accommodates a quite limited range of actions and excludes some mundane and habitual actions. Among these we can consider those "immersed" actions that involve what Dreyfus defined as "mindless coping" and of which driving the car is a classic example.[83] When I am engaged in one of these actions I display just a basic self-awareness. I know what I am doing and I am aware of being the agent who is doing it. I might not know the details of how I am doing what I am doing. Marcel[84] has argued that what is distinctive of the basic self-awareness associated with this action is the fact that one's self is not an object of awareness at all. As he points out, in our activities, our consciousness is immersed in our projects. If we attempt to turn our reflective regard from our projects to the structure of consciousness, or the self, we alter our intentional structure, and "the self who had been immersed in those projects is now

[83] Dreyfus, (2002).
[84] Marcel, 2003; Gallagher and Marcel, 1999.

abstracted from them".[85] Since in these activities my objective self-conception plays no role, it is not surprising that Korsgaard's model does not consider them as actions. One's objective self-conception is necessary for that "deeper" sense of ownership that Korsgaard ascribes to human agency. The limitations of this account of self-consciousness are particularly evident if we consider the following passage:

> When we move voluntarily, we move consciously. But this is not to say we are conscious that we are moving. Much of the time when we move nothing is further from our minds than the fact that we are moving. But of course this does not mean that we move unconsciously, like sleepwalkers. It is crucial, in thinking about these matters, not to confuse being engaged in a conscious activity with being conscious of an activity. Perhaps such a confusion lies behind Descartes' bizarre idea that nonhuman animals are unconscious. In the direct, practical sense, an adult hunting animal which is, say, stalking her prey, knows exactly what she is doing. But it would be odd to say that she is aware of what she is doing or that she knows anything about it. What she is aware of is her environment,the smell of her prey, the grass bending quietly under her feet. The consciousness that is inherent in psychic activities should not be understood as an inner observing of those activities, a theoretic state. An animal's consciousness can be entirely practical.[86]

Here Korsgaard draws a distinction between being engaged in a conscious activity and being conscious of an activity. One could be tempted to think that thanks to the introduction of this distinction Korsgaard can after all account for our mundane and immersed actions, since they seem exactly the kind of conscious activity I simply engage in. Nevertheless, significantly, Korsgaard links being engaged in a conscious activity to the behaviour of animals and talks of "voluntary movement" rather than action. The choice of words and the example used here are significant because they make clear that Korsgaard is not talking about the self-consciousness that is distinctively human and necessary for full-fledged action.

To sum up, I believe that both Korsgaard's and Velleman's models display this mischaracterization of self-consciousness: they assume that one's objective self-

[85] Gallagher and Marcel, 1999, p.289.
[86] Korsgaard, 1989, p. 118.

conception is necessary both to characterize the specificity of human agency and to link it to rationality. I argued that this picture of self-consciousness is not realistic and that, moreover, the reference to one's objective self-conception overlooks the specificity of human rational agency. This view is the very ground for those undue restrictions of the range of actions that we have seen in previous chapters associated with both views of reflective endorsement.

3.4. Conclusions

In this chapter I tried to understand the nature of the narrative self by first establishing its metaphysical commitments. I showed how Velleman conceives the self as a narrative self-*representation*, thus ascribing it a fundamental role in the agent's psychology while avoiding some substantial metaphysical ambiguities contained in Dennett's original view. This allows for the possibility that the notion of narrative self so conceived is neutral about the metaphysics of the subject whose psychology it helps describe. However, this left the task of establishing whether the self so described really is just an illusion, as Dennett claims. In order to answer that, one needs to show that one's reflective representation does in fact play the role Velleman attributes it in the agent's cognitive economy. In order to affirm with Velleman that the narrative self, although fictional, like in Dennett's description, is also *factual,* we need to confirm the role of one's self-conception in coordinating and guiding action.

While Velleman appeals to substantial work done in psychology to support narrative approaches, I argued that we cannot in fact get a definite answer to the question from this work. I believe that this is not just a matter of currently lacking empirical evidence: it is due to constitutional reasons inherent to the nature of the self-representation that narrative accounts assume. Empirical work in agentive awareness that builds a top-down approach, in which the central role is played by high-level, comprehensive representations, assumes the pervasiveness of misinterpretations, self-deception and wishful thinking.

While this is not much of a concern for narrative approaches seeking to define the conditions for agentive awareness, it is a problem for Velleman's causal model.

Finally, in my final section, I considered how this discussion bears on the thesis, which we find both in Velleman and Korsgaard, that there are two levels of selfhood, one of which is essential for agency. I questioned this thesis and argued that the level of self-consciousness described as distinctively human in both models requires a high level of complexity. I claimed that this is an unnecessary and unrealistic way to conceive of self-consciousness and that it is at the origin of many of the problems for these models that I have highlighted in previous chapters.

Chapter 4

Self-governance and Control

So far in this work, I have raised doubts about the cogency of the idea of self-representation that both models of reflective endorsement assume to be guiding behaviour in action. Now I want to focus on the notion of guidance and control they deploy. This will lead me to consider what it is for behaviour to be guided by one's self-representation according to Korsgaard and Velleman.

Models of reflective endorsement share the idea that human agency exhibits the distinctive feature of being self-controlled, self-governed and, in the Kantian sense of the term, characteristic of both Velleman and Korsgaard, autonomous. In both models self-governance is taken to amount to control over one's motives, in light of one's self-representation. Moreover, for both philosophers, for an agent to act is for her to cause behaviour in the right way: a way that allows for rational control. And we have seen that, for both authors, this means that actions are behaviour guided by one's self-representation. This means that, in both views, some psychological elements – drives, beliefs and desires for Velleman, and incentives and principles for Korsgaard – must stand in certain relations so that the behaviour they cause can count as action. One's self-representation is precisely what *governs* the relations of these different psychological elements: because of this, in both accounts, one's self-representation plays a crucial role in structuring these elements and achieving the unity of agency, thus determining behaviour as one's *own*.

In this chapter I will examine how this notion of control and self-governance is supposed to be specific of human actions, and the extent to which behaviour being controlled by one's self-representation amounts to its being self-governed. As we have seen, Velleman and Korsgaard consider two things necessary for agency: firstly, the behaviour should have the right psychological structure and, secondly, it should have the right relation to consciousness. The notion of conscious control, which is central to both views, then, is necessary to explain both the agentive awareness that goes with full-blown agency, and the characteristics of the causality, and psychological structure associated with actions.

In §1 I will show how one's control over one's actions relates to one's experience of agency in these models of reflective endorsement. I will argue that Korsgaard's model requires a very restrictive notion of control which mirrors the restrictions of her conception of agency criticized in chapter 1. As for Velleman, I will expand on the remarks I made already in chapter 2 and show that his conception of control is vague. I conclude that these problems contribute to making both views unsatisfactory.

In §2 I examine their notion of control in relation to the psychological structure and mechanisms necessary for agency. I do so by presenting what I consider a problematic case for reflective endorsement views of agency: anorexia nervosa. I criticize Korsgaard's account, which sees anorexic behaviour as akin to the behaviour of what she calls the 'tyrannic soul'. As for Velleman, I argue again that his notion of control remains vague and therefore unhelpful to provide a convincing understanding of what self-governed behaviour amount to. Since these views are unsatisfactory, I want to confront them with a widely influential view based on dual-process theories in cognition that sees self-control as a system 2 process. In philosophy, this view has been developed by Neil Levy and I will therefore consider whether it is possible to replace or integrate Korsgaard's and Velleman's notions of self-control with his. One might hope to solve the problems I highlighted in Korsgaard's and Velleman's views by recourse to something like Levy's account of self-control. I argue that this possibility is not open for Korsgaard:

her normative requirements for a conception of agency are inherently incompatible with the purely procedural notion of control we find in Levy. On the other hand, I argue that Velleman's conception of self-control fits better than Korsgaard's with views developed in cognitive psychology, which are close to Levy's approach and I point at the elements in Velleman's theory that need further development.

In §3 I present a last concern that we can have with theories that associate control with self-governance: namely, the extent to which the idea of an action's being self-governed really can correspond to its being guided or controlled by one's self-representation.

This chapter allows me to examine the notion of control in place in both of these theories and show its consequences, which are connected with the difficulties I discussed in previous chapters and that made these models of agency unsatisfactory.
The discussion in these three sections of agentive awareness, of the psychology underlying self-control and of the identification between self-control and self-governance shows that the notion of control these theories employ is problematic.

4.1. Control and Agentive Awareness

In what follows, I present some cases that can all be considered deviant from the normal control associated with robust agency in some sense, although in rather different ways. These cases have been widely discussed in philosophy and psychology to cast light on the notion of control. Through the discussion of failures one can hope to learn about the proper functioning of mechanisms involved in action control and also about the relation between control and experience, given that these cases present different breakdowns of the agent's agentive awareness. In examining these cases I will consider in what ways they present subjects with a diminished sense of control and I will say that we can understand these cases as presenting two levels of experience. This raises

the question about which of these two levels of experience is relevant for reflective endorsement views.[1]

Pathological Cases[2]

Anarchic Hand Syndrome: patients find their hand performing complex and goal-directed movements. These movements cannot be voluntarily interrupted or inhibited directly and they might interfere with the actions carried out by the other (healthy) hand. Della Sala characterizes these movements as "well executed but unintended".[3] To their dismay, patients perform movements resulting in picking up leftovers from someone else's plate, or they report that, in choosing TV channels, ``no sooner had the right hand selected one station the left hand would press another button'' and they resort to using the other (healthy) hand in order to stop the wayward behaviour, sometimes hitting it violently in anger and frustration.

---

[1] Given this agenda, it is important to stress that the list that follows does not aim at being exhaustive, but only to provide material sufficient and relevant to introduce this discussion.

[2] Most of the following are usually classified as psychopathologies affecting the frontal lobes (Joseph, 1999). The frontal lobes can be subdivided into some major functional neuroanatomical domains which, when injured, surgically destroyed or affected in their activity and volume seem associated with pathologies like Utilization Behaviour (left or right unilateral, or bilateral focal frontal lesions), depression, "blunted schizophrenia and aphasic perseveration (left lateral, or bilateral frontal and striatal lesions); mania (right frontal and bilateral disturbances); catatonia (medial frontal lesions); perseveration, obsessive compulsion (orbital frontal and frontal-striatal disturbances); Anarchic Hand (medial frontal cortex, in particular the supplementary motor area (SMA) but also the corpus collosum) etc.
I decided to leave out further references to these neuroanatomical details mainly because they do not play any central role in my discussion but also because I wanted to avoid committing to specific hypotheses in a field that recognizes substantive ambiguities: these subdivisions hold substantial interconnections and abnormalities are not always discrete and localized. This means that these associations are not always clear-cut and might involve abnormalities in different regions of the brain. In some cases, like perseveration, the associations with the frontal lobes is controversial (Schindler et al., 1984), in others, with particular reference to the Anarchic Hand syndrome, it had been noted that "the anatomical and neurophysiological specifications add little to our understanding of the mechanisms giving rise to anarchic hand. Furthermore, the lesions shown by patients are normally too large to allow any fine-grained anatomo-functional correlative speculation." (Marchetti and Della Sala, 1998, p.198). In others, like OCD, despite some agreement in the involvement of deficits of the executive functions of frontal lobes (Pujol *et al.*, 2004), matters are complicated by the relations with genetic, behavioral and environmental factors to trigger the disorder.
In light of these considerations, then, the present chapter will not offer a discussion of the neuroscience behind the mechanisms underlying control. In many contexts, this is one of the main topics of interest in the debate over these cases, but it is not relevant for the aims of this work.

[3] Marchetti and Della Sala, 1998, p. 196.

Global Automatisms: somnambulism, fugue or trance states. The term "global" refers to a global disruption of consciousness. The behaviour produced is complex, goal-directed but the awareness is very limited, with most patients tending to be amnestic about what they have done.[4]

Utilization Behaviour: patients automatically use objects that are within their vision and reach. Although they use the object in a manner that is instrumentally correct (for example, presented with a pair of glasses they would pick them up and put them on), they cannot stop themselves performing these actions, even when asked not to and even if they consented to the request: if several pairs of glasses are presented to him, the patient would end up wearing them all.[5] As Lehrmitte puts it, it is as if "the presentation of objects implies the order to grasp and use them".[6] Patients do not appear to understand that they are doing anything inappropriate, even though they can repeat the instruction to desist. If asked, they will calmly and correctly say what they are doing while fabricating reasons why they did it or just saying that they felt like it.[7]

Obsessive-compulsive behaviour:[8] compulsive behaviour comprises urges to do something over and over again and leads to patterns of behaviour that the person feels she has no power to overcome. Patients of OCD experience intrusive and repeated thoughts, feelings or sensations (obsessions) that generate a sense of anxiety and distress; they adopt patterns of behaviour (compulsions) they feel driven to usually in order to reduce their anxiety and distress.

---

[4] As defined in Levy and Bayne, 2004, p. 210.
[5] Lhermitte,1983, in Pacherie, 2007b.
[6] Lhermitte,1983, p. 237.
[7] Pacherie, 2007b, p. 215.
[8] Here I refer to both cases of OCD (Obsessive-Compulsive Disorder) and OCDP (Obsessive-compulsive personality disorder). The relations between these two disorders gives rise to a large debate and is a controversial topic, with several theorists pointing at the differences between the two disorders (compulsions in OCD seems to revolve around rituals, whereas OCDP is mostly associated with perfectionism; OCD is ego-dystonic, whereas OCDP seems rather ego-syntonic), while others stress the co-morbidity between the two can constitute a subtype of OCD (Garyfallos, *et al.,* 2010). I do not mean to enter this debate or make any substantial claim by considering the case of compulsion in general.

Impulse-Control, and Conduct Disorders: there is a vast array of disorders to consider here,[9] but they all result in behaviour that violates the rights of others (e.g. aggression, destruction of property) and bring the patients in conflict with societal norms. Among these we find, for example, intermittent explosive disorder (outbursts that result in physical or verbal aggression). Patients who suffer from these disorders experience recurrent unpremeditated outbursts of anger that are not due to other disorders (e.g. depression, antisocial personality disorder etc.) and are egodystonic (the patient expresses distress, remorse etc.).[10]

Dissociations between verbally acknowledged intentions and actions: patients express verbally the intention to act, but they fail to carry out the intention. Otherwise they perform some act they affirm they intend not to. The intention appears to be present, but ineffectively so. For example, when asked to measure a piece of string in order to cut it later, the patient agrees with carrying out the task but immediately starts to cut it while remarking "Yes, I know I am not to cut it".[11]

Non-pathological Cases

Slips of Action: they happen to most of us and can involve something quite basic like saying the wrong word, as well as complex coherent and well-coordinated tasks (as in James's famous example of going upstairs to change for dinner and finding oneself performing one's bed routine). When attention is drawn to the behaviour, it is immediately recognized as not intended. The causes of such slips are extremely varied: Freud famously interpreted these errors of speech and action (as in the case of the broken inkstand or of the president starting the meeting by announcing its closure) as revealing one's desires and

[9] And they vary in the way that they are classified. For example, Kleptomania and Pyromania, which were listed as distinct disorders in DSM-5, have been removed in its revision due to insufficient evidence to retain them as distinct disorders. Kleptomania is now considered "compulsive stealing", a symptom of depressive disorder or impulse-control disorder. Similarly, pyromania has been considered as "fire-starting" within Conduct Disorders.

[10] The discussion of ego-syntonic and dystonic disorders will come into play later, for the moment it is sufficient to notice that not all these disorders need to be in conflict with one's self-image and can, for example in the case of pyromania, be felt as acceptable to one's needs and goals.

[11] Monsell, 1996, p.119.

intentions. But most of these tend to be considered cases of absent-mindedness.[12] Among slips I include *Capture Errors*:[13] in a task composed of multiple steps, some stronger habit takes over at the point in which one was to depart from a familiar course. So driving home but planning to stop at the groceries one finds oneself driving through as always, maybe realizing just once at home that one failed to stop. The label alludes to the fact that one's attention is captured by something outside the task, resulting in capture of behaviour by some inappropriate (habitual) activity.

Lost Intentions: a familiar case which has already been discussed already in chapter 2 is the one of someone stopping at the kitchen door asking "what am I doing here?".

Dispositional and Emotional Reactions: sometimes one wants to do something most of all but psychological factors get in the way. Here are some examples presented by Jeanette Kenneth:

> I break down while giving the eulogy at the funeral of a loved one and am unable to complete the tribute I so wanted to make. I want to ask Arnie out for dinner but I'm overcome by shyness as I look into his ice-blue eyes and I blush and stammer incoherently. I fail to suppress a sneer that betrays my contempt for the executive to whom I'm making an important presentation. I tense up and botch the dive that would have earned me Olympic selection.[14]

Weakness of the Will: Weak-willed agents fail to control themselves. In our everyday life we consider most of these cases as occasions in which we "give in" to some desire and fail to stick to our resolution.[15] So, I might go ahead and have another slice of chocolate cake, even though I am on a diet, or I resolve to

---

[12] "Those who have stepped into their baths wearing some garment, or struggled to open a friend's front door with their own keys, or switched on the light as they left the room in the daytime, or attempt to pour a second kettle of water into a pot of freshly made tea, or turned off the television set when they meant to extinguish gas fire, or said "Thank you" to a stamp machine, will recognize the species. Our daily lives are strewn with such trifling and usually inconsequential blunders", (Reason, 1984, p. 517).

[13] Monsell, 1996, p. 110.

[14] Kennett, 2013, p. 151.

[15] A discussion of weakness of the will in detail is tangential to the aims of this chapter and I will not enter controversies in distinguishing akrasia from weakness of the will (Holton, 2009), and whether we are to treat weakness of the will as desire-based or judgement-based. I will touch upon this when discussing Levy's theory of self-control later on.

leave the pub early, to function better at the important work meeting I have tomorrow, but I end up staying for another drink, etc.

I will first consider how these cases can be analysed to understand self-control and how they relate to agentive awareness. Then I will connect these considerations to Velleman's and Korsgaard's view.

In order to conduct this analysis of the relationship between these cases of failure of control and agentive awareness it will be useful to employ Shaun Gallagher's distinction between sense of agency (SA) and sense of body-ownership (SO), and, within the sense of agency, between two levels: SA(1) and SA(2).[16] The sense of body-ownership (SO) for a movement is the sense that *I* am the one who is undergoing the movement, I am the subject that is moving. This SO is present even in involuntary movement inasmuch as it is my body that is moving. SA(1) is a basic and minimal self-awareness that I am the one causing or generating the action, that I am the author of my movement. In the normal experience of voluntary movements SA(1) and SO are indistinguishable However, I have a different experience when I dive than when I have been pushed to fall: we can see that, in the latter case, it is not *I* who has caused the movement, although I still retain the sense that it is my body that is moving. SO is thus defined by Gallagher as a pre-reflective experience that is taken to originate from afferent sensory-feedback (visual and proprioceptive/kinaesthetic information that tells me that I'm moving), but there are no initial (efferent) motor commands that generate the movement.[17] SA(1) seems to result from efferent processes,[18] and is pre-reflective, which means that it does not require introspection, and that neither metacognitive

[16] These distinctions are employed most notably by Gallagher, 2000; 2012; Marcel, 2003; Tsakiris *et. al*, 2007.

[17] See Gallagher, 2010, Tsakiris *et al*., 2007, for evidence that the sense of agency originates in neural processes responsible for the motor aspects of action and also Haggard, 2005; Lau, Rogers, Haggard, & Passingham, 2004. My discussion of agentive awareness will not take on explaining the mechanisms that link the sense of agency to the motor commands sent to the muscles and the accompanying efference copy that is internally processed. Equally, I will not focus on the predictive models of the motor system that these considerations support. See Haggard, 2005, and Tsakiris and Haggard, 2005, for the role of efferent binding in generating the experience of agency.

nor explicit perceptual monitoring of one's doing is necessary: my agency is not normally something that I attend to or something of which I am explicitly aware. Because of this, SA(1) is a conscious experience but remains a recessive aspect of consciousness. This kind of basic experience of agency can be distinguished from a higher-order reflective phenomenon SA(2). I have SA(2) for my actions because I have a properly ordered set of interpretations of my doing in terms of the beliefs, desires and intentions involved and these mental states would normally explain my actions. Deliberation and planning often contribute to one's SA(2). This means that "whether I count something as my action thus depends upon whether I take myself to have beliefs and desires of the sort that would rationalize its occurrence in me".[19]

SA(2) and SA (1) have different intentional contents. For SA(1) to be an experience with an intentional content conscious attitudes and judgements need not be involved. Rather, it seems that there is some pre-reflective awareness of the movement: I have a basic feeling of doing that comprises both the experience that *I am the author* of the movement and the experience of *what* I am doing, of the effects of my behaviour. It is in light of the latter that we can understand a kind of pre-reflective monitoring that allows for the quick correction of my movements and requires no conscious judgement. For example, when I reach to grasp a cup I do not need to consciously verify the direction of my movement or the shape of my grasp. However, there is also substantial evidence that, even at this basic level, my movements are directed in a way that discriminates between different intentions: so for example, my grasp will automatically adapt to the objects (the size, shape and orientation of the cup for example) and my goals (I will grasp things differently if I am reaching for the cup in order to drink or to throw it, for example).[20] On the other hand, SA(2) involves conceptual thought, conscious judgements and monitoring in terms of one's mental attitudes. Notably, in SA(2) there is a different sense in which one has an experience of *what* one is doing, one that is

[18]Haggard, 2005. Note that efferent processes are usually called 'motor control processes' in the literature. I am avoiding the terminology since the subject of 'control' more generally is the topic that this chapter wants to address.
[19] See Gallagher, 2012, p. 18, citing Graham and Stephens, 1994, p. 102.
[20] Pacherie, 2011. See also Gallagher, 2012, p. 24.

linked to his sense of *why* one is doing it: one's actions are thus understood under descriptions that explain the action in terms of the agent's reasons.

With these distinctions in place, I will consider which dimensions of agentive awareness are involved in the different cases of failures of control I presented. In the light of this analysis I will then try to establish which aspects of agentive awareness seem to be relevant for the kind of control that Velleman and Korsgaard require for full agency.

The loss of SA(1) is defining of phenomena like the Anarchic Hand Syndrome. Unlike cases of hemisomatoagnosia or Alien Hand, in which patients deny ownership of their own hand (hence displaying absence of SO for a particular part of their body), Anarchic Hand patients do not disown the hand that performed a certain movement: they disown the movement itself. In fact, the distress experienced by the patient is relative to the fact that the hand performing the disowned behaviour is theirs: one has lost the capacity to implement his intentions when it comes to his anarchic hand, and in this sense it clear that they also have no SA(2).

Cases of Utilization Behaviour, while displaying similar impairments in the mechanisms of action production,[21] present a very different phenomenology because patients do not disown their actions and there is a distinctive lack of concern, in striking contrast with the discomfort and sometime outright terror experienced for anarchic hand acts.[22] *Prima facie* it would seem that, while the

[21] Marcel notes that "Anarchic hand has a experiential aspect and a component of action control; the former is frequently confused with Alien Hand and the latter with Utilization Behaviour". (Marcel, 2003, p. 76).

[22] Marcel discusses this asymmetry in regards to a patient who suffered of both conditions. See ibid., p. 78. The relationship and connections between Anarchic and Syndrome and Utiliazation Behaviour is complex and controversial (Pacherie, 2007b). In both conditions, although it is clear that at some level the mechanisms of action production are impaired, it is clear that the low-level motor programs are intact: the movements these patients perform are appropriate to the specific object in the environment at which they are directed, even though they may not be appropriate to the wider context. However, it is controversial that behaviour is in both cases environmentally and stimulus-driven. Marcel (2003), for example, considers that this is not the case for Anarchic Hand syndrome. Pacherie notes that patients suffering of Utilization Behaviour seem also to have lost the capacity to generate and act on endogenous intentions, whereas Anarchic Hand Syndrome patients have not lost this capacity. She cites Lhermitte's study that shows that sufferers of Utilization Behaviour are not only powerless in the face of influences from the outside world, but they exhibit inertia and apathy when there is no external stimulation. In Anarchic Hand Syndrome the capacity for full-fledged intentional

executive mechanisms producing action are malfunctioning and movements are triggered by environmental stimuli with no possibility for the agent to stop them or inhibit them, patients experience both SA(1) and SA(2). However, even allowing that patients of Utilization Behaviour still have the sheer phenomenal experience of agency associated with SA(1), their SA(2) seems impaired. Although these patients never exhibit surprise or perplexity at their own behavior, their confabulations are not elaborate and they justify their actions with claims such as "I thought I had to do it" or "I thought you wanted me to use them"[23] and if pressed about their reasons they might admit that they do not know why they did what they did or say that they just felt like it.[24]

Cases of Impulse Control and dissociation between behaviour and verbally acknowledged intentions, seem to present us agents who feel driven to do certain things. In these cases it is not SA(1) that is absent: these agents retain the basic awareness that it is them who are causing certain effects and they are not just experiencing SO. They have a basic feeling of doing but they lack SA(2) and, to their despair, they perform some behaviour they disavow. The case of obsessive compulsive behaviour is more ambiguous[25] but I believe that in general we can understand it along similar lines: when OCD patients feel *compelled* to behave in certain ways *in spite of themselves,* they may not conceive of themselves as "the agents". In this sense their SA(2) is diminished, if not absent. It seems that these agents lack a sense of control inasmuch as they do not experience these compulsive actions as ones they initiate and that they have the power to stop.

---

action is retained, as it is clear when the healthy hand tries to stop the other's movements. Pacherie's hypothesis is then that this capacity of implementing one's own intentions is impaired in Utilization Behaviour and that "having lost the capacity to will their actions, they have also lost the capacity to experience them as un- willed" (2007b, p. 216). The possibility to attribute SA(1) to Anarchic Hand patients is also controversial. I have been urging that SA(1) is lacking, following Marcel (2003). Others, like Pacherie, consider that the fact that the low-level motor programs are intact allows for a basic self-awareness. Regardless of how to settle the score on this issue, for the purposes of this chapter it is important to notice that controversies regarding SA(1) do not invalidate the consideration that SA(2) is absent in both cases (despite Utilization Behaviour retaining a diminished sense of it).

[23] Pacherie, 2007b, p. 215.

[24] Marcel, 2003, p. 77.

[25] For example, when the behaviour people feel compelled to is ego-syntonic (as it seems in OCDP), it might seem that patients retain a meaningful dimension of SA2.

Among non-pathological cases some are straightforward: lost intentions present a case of behaviour for which we have a minimal agentive awareness in terms of SA(1) but we do not experience SA(2). So, for example, when standing at the kitchen door wondering why I got there, I know that I acted and walked to the door, but I do not have a sense of what my reasons for this action are. I might retain a sense that I did head to the kitchen with some aim, but I lost the SA(2) that would allow me to explain my intentions.

Other cases, like those of slips of action, seem to vary substantially. In James's classic story of the man entering his room to get ready to go out and finding himself performing his bed routine we have an example of how SA(1) can be retained: at his wife's bewilderment when finding him at getting into bed, he replies "I did it automatically", thus not disowning the action in the basic sense that would be require by the absence of SA(1). It is SA(2) that is definitely missing in these cases. By definition,[26] they slip through one's intentional control. However, in some other types of cases, they can present a different phenomenology, which also marks them unintentional. I might, in fact, fail to acknowledge that I acted at all: I might not realize at all that I have knocked down the inkstand, even if my movement has been driven by my unacknowledged beliefs and desires and it was not a mere accident. While I might, at the moment, retain some experience of my moving in a certain direction, I might fail to understand or experience its result (in a way that seems to rule out the intentional component of my sense of what effects my action brings about, which, as we have seen, is integral to SA(1)). I might come back later and judge that I have broken the inkstand without any experience from the inside that I did. Even if the fragile and (probably short-lived) experience of my movement could be retrieved in memory,[27] this is not the experience of causing certain results.

[26] Here I am employing Velleman's own definition. See Velleman, 2000, pp. 3-5.

[27] Memory does complicate matters here because it seems that it is particularly fragile when it comes to actions for which SA(1) but not SA(2) was present: so I can move absent-mindedly my car keys, or my phone and I might have some vague recollection of having done it (I can start tracing back my steps to find the items I moved and cannot find anymore). I say that memory complicates matters because it seems that I can sometimes, for example, move absent-mindedly and then struggle to find something while remembering my intentions and the

The category of Emotional Reactions needs to be restricted to those that can count as something I do, since some (like blushing, fainting etc.) seem to be more physical reactions like reflexes than actions at all. It seems that generally we can retain our SA(1); I know that I am smiling even if I do so unintentionally (I might even know that I am smiling inappropriately for the situation, like people who cannot stop grinning at funerals). However, it seems possible that I might actually have no experience, not even in SA(1) sense, of my reaction and come to know that I actually am smiling because someone else points it out to me.

In cases of weakness of the will we seem to retain both SA(1) and SA(2). What seems to cause the experience of one's agency being diminished is a conflict experienced at SA(2) level, and the mismatch between one's weak-willed action and one's resolution. Indeed, when one acts in a weak-willed way, he does not reject the accusation of having behaved out of lust, greed or gluttony, and might even consider that his action exposes his character and think something like: "Here I go again".

The discussion in this section has tried to highlight how cases that present agents as failing to control their action actually vary substantially in the way the agents' experience their agency as diminished. In some pathological cases the agent fails to experience SA(1) for his action altogether, and this seems connected with impairments of the efferent processes that originate the experience at the motor level. When SA(2) is absent or diminished the agent experiences a loss of his intentional control over actions of which he is aware of being the author. Since this is the level that provides action explanation, it is SA(2) that seems interesting from the standpoint of reflective endorsement theories. It is at this level that we find conscious judgements and monitoring in terms of one's mental attitudes in a way that allows for rational control. We have already seen in the previous chapter that the kind of ownership that is entailed by the very fact that I am the agent of some activity is not the sense in which reflective endorsement theories demand actions to be owned. Because of this it is not sufficient for actions to be owned in the minimal sense associated

---

reasons that I moved it, thus having moved it with what might seem SA(2). E.g. I might

with SA(1), in order to count as self-governed and controlled. We can thus understand different accounts of reflective endorsement as providing different views of how we are to understand SA(2) and what it is to disavow one's action.

In the next section I will consider Korsgaard's position on agentive awareness and then Velleman's. My aim is to test whether in reflective endorsement theories of action it is the self-control associated with SA(2) that matters, as SA(1) is associated with basic functions of motor control connected with a kind of agency that does not present the specifics of robust agency. If this view corresponds to Korsgaard's and Velleman's, we would have a clear understanding of the relationship between self-governance and agentive awareness because self-governed actions would be associated with SA(2), while SA(1) could remain a feature of mere purposive activities. In what follows, I will argue that Korsgaard and Velleman do not hold this position: Korsgaard's view requires a more restrictive understanding of SA(2) and Velleman's view is ambiguous about the level of agentive awareness it requires.

4.1.1 Korsgaard's conception of agentive awareness

The idea that SA(1) is a feature of mere activities, whereas self-governed actions display SA(2), might seem confirmed in light of the following considerations: when Korsgaard says that in acting, "it is as if there is something above and beyond" your particular desires, the experience of agency that she is describing is not the thin and recessive experience that I have characterized as SA(1). Korsgaard's view requires one's practical identity to play an essential role for the agent's endorsement of his action and it cannot be associated with the pre-reflective and purely synchronic character of SA(1).

However, the conception of control necessary for self-governance cannot just be associated with SA(2): it requires a restriction at this level of agentive awareness. For example, Korsgaard describes wanton agency as defective and

---

remember that I moved my jumper, "to make sure I would pack it".

essentially lacking the features that can qualify it as robust agency. A wanton is someone who has lost the guiding thread that makes "his life worth living and his projects worth undertaking".[28] This means that wanton agency exemplifies a loss of control, according to Korsgaard's view, in which humans cannot operate without plans and overarching goals. Arguably, though, wantons still experience their actions as intended and they can provide explanations in terms of their beliefs and desires in a way that might seemingly rationalize their actions. When Korsgaard claims that they are not self-controlled then, she seems to require a rather different sense of control than the one associated with SA(2).

Before assessing Korsgaard's restrictive understanding of the agentive awareness associated with self-control, I briefly want to point at some difficulties with the way Korsgaard describes the experience of being a wanton. She claims that an agent "might think of herself as the slave of passions, and then she will be a wanton".[29] I believe that there is a way in which this claim is wrong: the phenomenology that Korsgaard describes here seems rather one of an addict or someone suffering from compulsive behaviour. A wanton is more likely to think of herself as someone who does what she feels like, and does not, after all, think of herself as someone losing control over her actions. In this sense Korsgaard would be making a wrong claim about the subject's self-conception. For Korsgaard, wantons are not just weak-willed agents because weak-willed agents endorse principles and fail to respect their commitment to them, whereas wantons do not commit to these principles in the first place. Korsgaard's description of wantons as slaves of passion seems then to be a consideration concerning the structure of the agent's will that one can draw from an external standpoint, rather than one about the agent's experience. I have discussed in chapter 1 how Korsgaard is sometimes ambiguous in distinguishing between the two, and, in particular, how the notion of identification is problematic because it seems linked to considerations that do not pertain to the agent's phenomenology, while it also seems to refer to the agent's experience.

[28] Korsgaard, 1996, p. 102.
[29] Ibid., p. 101.

These specific concerns about the description of wanton behaviour do not alter the fact that, in Korsgaard's picture, it is impossible to associate SA(2) with self-governance. SA(2), as I have presented it, requires the subject to consider his own attitudes, his desires and intentions: this is something different from thinking these attitudes as falling under a principle of choice, as Korsgaard urges. For Korsgaard the experience required for conscious control is one of identification with the principle she considers defining of our nature as agents: the categorical imperative. While the two are not incompatible, it is clear that this experience is a restriction on the more general SA(2) that we can associate with action explanation in general. It is clear, then, that while none of the agents described in the examples of control failures discussed above can count as self-governed in Korsgaard's model, the generic understanding of SA(2) does not lead to identify the kind of control that Korsgaard considers defining of human agency. It seems, then, that we cannot associate the distinction between kinds of agency with that SA(1) and SA(2), and self-control can, at the most, be a very restricted sense of SA(2).

4.1.2 Velleman's conception of agentive awareness

Velleman himself discusses some of the various cases of failure of control I presented above, and he does so in a way that seems to confirm the association of SA(1) with mere activities and SA(2) with self-governed actions.

For example, he suggests that both lost intentions and emotional reactions, which at the most exhibit SA(1), can be turned into instances of robust agency through the control we can associate with SA(2). Velleman claims that "the smile that springs spontaneously from your emotion of surprise isn't aimed at any result, but it, too can be transformed into a full-blooded action if it is brought under your conscious control".[30] This seems to require the subject's access to those intentions, beliefs and desires that also explain his actions to align behaviour with the subject's aims.

[30] Velleman, 2000, p. 91.

Nevertheless, Velleman's view on agentive awareness becomes more ambiguous if we consider his treatment of slips of action. In discussing them Velleman seems to suggest that it is the psychological structure that matters for agency, rather than agentive awareness. Only some slips are revealing of one's own intentions, whereas others are simply cases in which some routine takes over or an agent behaves absent-mindedly. These cases present lapses of attention, much like the "capture errors" described above. In this, they are different from the previously discussed example of Freud's breaking of the inkstand, where we can see the behaviour as actually selected in response to certain beliefs and desires of the agent: the desire to have a new inkstand and the belief that his sister would buy him a new one. When I go in the kitchen to open the window, and I end up opening the cupboard, I might not have any particular underlying intentions, as for example I would have if I opened the cupboard guided by some subconscious desire to check whether my flatmate is stealing my food. This difference is one that pertains to the underlying psychological structure of my motives rather than agentive awareness and it is important to underline that the cases that fundamentally interest Velleman are those in which the belief/desire structure is operative in leading the agent to perform some behaviour, but the cognitive drive for self-understanding does not play an active function in producing the behaviour. Only those slips resulting in some sense from the beliefs and desires of the agent count as that form of minimal agency that we can attribute to purposive activities (and that Velleman considered Davidson's account to describe). This means that only *some* of the behaviour that can be associated with SA(1) can be credited with having this structure and so it remains uncertain how the model handles and defines the rest of the cases that do not hold this relationship to the agent's hidden intentions.

Moreover, I presented above the case of an agent failing to experience his movement, which for example results in breaking the inkstand, as an action. In this case, based on Velleman's discussion, there is a belief/desire structure that makes the behaviour count as an instance of activity. However, since this minimal form of agency does not see the cognitive drive play a role, it does not

count as full action. Now, since in this example I do not experience my movement as an action at all, this basic activity comes with no SA(1). Because of this mismatch between SA(1) and the basic form of activity, there seems to be no parallel, in Velleman's model, between kinds of agency and the two levels of experience.

This picture is complicated when we consider how, in some of the cases, the control Velleman associates with full-blown agency seems to require a restriction on SA(2). For example, consider Velleman's treatment of cases of weakness of the will: these agents are presented as not exhibiting the control required for actions, because they fail to understand their behaviour at some level and thus their behaviour displays a failure of practical reasoning.[31] For Velleman, if someone regards himself as lazy (instead of laid-back) he must see his own behaviour as not entirely making sense. This seems to suggest that, like Korsgaard, Velleman considers that one's mere ability to understand one's intentions, as it is presented in SA(2), cannot suffice to describe the kind of control necessary for action. Unlike for Korsgaard, though, what is required for Velleman is that one sees these attitudes as coherent with one's self-conception.

However, the view that identifies the agentive awareness associated with self-control to this restrictive sense of SA(2) is problematic because it is in tension with Velleman's attempt to accommodate daily routines and ordinary activities within his picture of agency. We have seen already how the narrative model conceived a distinction between perceptual and deliberative reflection, so as to account for these mundane acts as instances of actual agency. We do not consider agents performing these kinds of activities as not in control of their actions, and a picture of agency that ruled these cases out would be severely limited. As we have seen, some of these activities, such as driving a car, seem to involve what Dreyfus describes as "spontaneous transparent coping"[32] and they present agents that are absorbed (immersed) in what they are doing in a way that excludes that they are reflecting upon their reasons and intentions to

[31] Velleman, 2007a, pp.202-203 and 246-252.
[32] Dreyfus, 2002.

do these things. In these cases I know what I am doing and I experience myself as the agent who is doing it. If agency requires a restrictive sense of SA(2), then it remains unclear how Velleman's model can accommodate these actions.

In order for Velleman to account for these activities, his view has to rely on the idea that conscious control and practical reasoning do not require the occurrence of actual thoughts about one's reasons and intentions. This means that accessibility, rather than actual access, is what matters for self-control and guidance in the narrative model. Since we can make sense of our behaviour in immersed actions, and we can access action explanations for them, it seems irrelevant that these activities constitutively require a minimal form of awareness. Once we understand SA(2) as awareness of agency that only requires accessibility, we can consider immersed actions as instances of self-controlled behaviour. I have already explored above the ambiguities and the difficulties that we find in the narrative model's treatment of the relationship between the phenomenology and this kind of accessibility. Here I simply need to stress that if we understand Velleman's narrative model in these terms, it sharply distinguishes between the access to action explanation necessary for full actions and the experience of agency. What the model does not provide is, therefore, an explanation of the relationship between self-governance and the experience of agency

Moreover, consider the famous case of someone's arguing with a friend and how this presents a case of diminished agency in Velleman's model. It is clear that for Velleman the example is presented as one that is not symptomatic of some pathology, as it would be if it was a case of intermittent explosive disorder, in which I suffer from some impairment in stopping my aggressive behaviour. Moreover Velleman's example is also not one in which the subject is entirely unaware of what he is doing: it is different from a case in which, for example, one would deny decisively, while shouting, that one is shouting at all. Velleman's case seems to present us with something done intentionally. In the argument with a friend, as it is presented by Velleman, it would seem too strong to say that the person experiences his shouting as something that he

does not actually do or something that he has no power to stop. It would also seem that the agent understands his action, and himself, under some description: he is angry, and acting accordingly. Despite this, Velleman claims that the behaviour does not count as fully acted and that the agent can claim that "it was anger speaking, not I". Velleman argues that this is the case because those explanations I find after my outburst do not function as reasons *in light of which* I behaved. One can understand this as meaning that, despite the fact that the agent understands his behaviour under some description, he misses a richer understanding of his behaviour in light of reasons connected with his narrative self-conception. If so, though, the view implies a demanding form of self-knowledge and it is also not clear that it pictures the case as one of failure of control. Moreover, if, as it seemed the case for the immersed actions discussed above, one associates self-control with accessibility, rather than with actual access, there is no reason to consider that this case of arguing with a friend is not an action: these motives were operative in my behaviour and I could make sense of them, and of my behaviour in light of them. Velleman thinks that this is not the case because these considerations, so to speak, *came too late*: my anger (and the reasons supporting it) is not the rationale for my behaviour and they are not the considerations in light of which I actually performed it. This position makes unclear the role of accessibility in Velleman's picture.

To sum up, in this section, I questioned whether, in defining the agentive awareness associated with the self-control necessary for robust agency, reflective endorsement views are uninterested in the basic feeling of doing that we can associate with SA(1). The kind of self-control that is pertinent to those models is not one that goes amiss when SA(1) is lost and is most likely to be associated with SA(2). On examination, however, it does not seem that we can associate SA(2) with the dimension of full agency that Velleman and Korsgaard wish to capture.

In Korsgaard, the kind of self-control necessary for action requires a very restricted conception of SA(2): self-control is linked to the identification with the categorical imperative, the principle governing the unity of agency and

identity. Korsgaard's model gave us a very limited range of behaviour that amounts to action, and this is mirrored by the kind of control experienced by the agents, which is also limited to a very specific sense of SA(2).

Velleman's conception of conscious control is more obscure. At times he seems to associate it with SA(2) (as in the case of lost intentions and emotional reactions); at other times SA(2) seems to comprise a richer understanding that relates to the agent's narrative self-conception. At other times his treatment of slips and immersed actions seem to indicate no connection between control and the experience of agency. Agentive awareness here seems to be understood in terms of possible access to action explanation. While one can argue about the plausibility of Velleman's notion of accessibility and the role that it plays in agency, it also remains a fact that Velleman's story is in need of some account of how to connect experience and awareness of agency.

Both Korsgaard and Velleman, then, fail to offer a satisfactory account of the relation between control and agentive awareness, which is presented in restrictive terms in Korsgaard and obscure ones in Velleman. In what follows I will argue that both theories also struggle with defining self-governance in light of the psychological structure it requires.

4.2. Self-control and Psychological Structure

As we have seen, both Korsgaard and Velleman consider that the right psychological structure is necessary for agency. The kind of control over one's motives essential for self-governance requires that one's practical identity, or one's self-narrative, structures the elements of one's psychology through reflective endorsement. In what follows I want to discuss the case of anorexia nervosa and the challenge it can present to reflective endorsement theories of agency.

According to the *International Classification of Diseases, ICD-10,* a central feature of anorexia nervosa is the 'deliberate weight loss'.[33] While some sufferers willingly undergo treatment, a substantial part of anorexics think that their eating habits are 'normal', and they prefer the pleasure and elation of fasting and exercise over other pleasures (eating) they openly disvalue. They endorse their anorexic behaviour and have a strong conviction about what to value:

> Fasting has been associated (and is still associated) with ideas of control over the chaotic passions of the body, and the person who is able to exert control over hunger, such a powerful physiological impulse, has often been presented as an example of moral integrity.[34]

Anorexics are committed to the value and *pursue of lightness,*[35] and see food restriction as exemplifying self-government, discipline, and the submission of bodily urges. Anorexic behaviour seems radically opposed to that of weak-willed agents and wantons: on the one hand, weak-willed agents fail at keeping their behaviour in line with their resolutions, and this failure makes the judgements and the commitments they form ineffective. On the other hand, wantons have no overarching value to provide unity and purpose to their actions: they just seem not to ever stand back to evaluate the ends their desires incline them to pursue. This seems to condition and diminish the sense in which they can resist some affective states (their food cravings, for example, or their sex urges etc.): it is because of this that Korsgaard considers that wantons are "slaves to their passions". They seem able to form plans about the paths to follow in order to accomplish their preferences, but it would be inappropriate to employ the term 'resolutions'. This is because they are able to develop means-end strategies and are capable of weighing current options, but they live in the present and are unable to stick to their resolutions. But a plan that could, at any time, and for any reason, be abandoned because it does not entail any real commitment that supports its choice and application, does not seem to be a "resolution". By contrast anorexics firmly endorse values that support their fasting, they stick to their resolutions and their decisions to

[33] World Health Organization, International Classification of Diseases, (Geneva: WHO, 1992), F10-19, as cited in Giordano, 2005, p. 94.
[34] Giordano, 2005, p. 127.
[35] Ibid., p.95.

restrict food intake also seem deliberate and reasoned. Consider how Nomi Arpaly rephrases Korsgaard's famous exemplification of the inner monologue of a rational and autonomous agent from the standpoint of someone who has anorexia nervosa:

> I see a piece of cake in the fridge and feel a desire to eat it. But I back up and bring that impulse into view and then I have a certain distance. Now the impulse does not dominate me and now I have a problem. Is this desire really a reason to act? I consider the action on its merits and decide that eating the cake is not worth the fat and the calories. I walk away from the fridge, feeling a sense of dignity.[36]

Anorexic behaviour is particularly interesting for the discussion on self-control because it presents several elements that we normally associate with self-governed agency: it is goal-directed and requires strong dedication and planning; it is not ego-dystonic, but rather it seems to be in line with the person's convictions and values to the point that some consider it as a life-style.[37] Anorexics actually seem to believe that their fasting enables them to get a grip on their lives. In terms of their experience, they seem to experience something similar to what Frankfurt described as cases of volitional necessity:

> There are occasions when a person realises that what he cares about matters to him not merely so much, but in such a way, that it is impossible for him to forbear from a certain course of action. It was presumably on such an occasion, for example, that Luther made his famous declaration: "Here I stand; I can do no other".[38]

Their experience is not one in which this impossibility is *felt* as their being unable to forbear because one is driven to act by some compulsion too powerful to be overcome. However, the fact that these patients appear to themselves to be in control of their desires and making value choices, contrasts sharply with our intuition (and the intuition of their families and friends) that these people are in fact at the mercy of their desires, of their obsessions and of their fear of losing control. We do not think that they are, after all, self-governed agents. How do we explain the intuition that these agents are less

[36] Arpaly, 2003, p.18.
[37] As is dramatically clear from several "pro-ana" websites that proliferated in the past years.

than autonomous and their behaviour lacks self-governance and control? In what follows I will examine how Korsgaard's and Velleman's models of agency explain the case of anorexia nervosa. I do so in order to show the difficulties that these cases present for their view and in general to the idea, present in both philosophers, of equating self-governance and self-control.

4.2.1 Korsgaard's model and anorexia

Despite the fact that anorexics experience their behaviour to be in line with their values and self-controlled, according to Korsgaard's model, anorexic behaviour is similar to what she describes as behaviour produced by a "tyrannic constitution".[39] For these agents there is one end which they are going to pursue no matter what, and effectively it rules over them. This end settles everything in advance prior to reflection. This means that their supposed endorsement does not allow for their conviction to enter the space of reasons at all.[40] People with anorexia nervosa maintain their capacity for planning and, for example, control their behaviour according to the social context, and this would credit them with considerable "instrumental intelligence";[41] nevertheless their illness determines what is worth doing for the sake of what. Korsgaard claims that "the tyrannized soul can never separate himself from one of his impulses, and so consolidates himself into a mere force of nature, an object, a thing".[42] Like the tyrannized soul's, the anorexic's behaviour *seems* unified but the appearance of unity it achieves is actually the result of one's obsession. For Korsgaard their obsession clashes with the categorical imperative, which is the principle that constitutively belongs to their human nature. In this sense, the tyrannic constitution is at the very end of the spectrum of integration and unity that Korsgaard considers as constitutive

---

[38] Frankfurt, 1971, p. 86.
[39] Korsgaard, 2009, pp.171-173.
[40] As I will point at later, this view seems to equate eating disorders with a type of addiction (Giordano, 2005, pp.76-80) and as such associates them with compulsion. There are a lot of critical assumptions here, some of which I will explore later on: first, one needs to establish that eating disorders are a form of addiction; then one needs to consider the extent to which compulsion defines addiction. There are substantial controversies on this last point, which is the reason why I resolved not to discuss cases of addiction in this chapter. While the view of addiction-as-compulsion is the dominant one, there have been important challenges to this idea. See later in my discussion of Levy, 2006.
[41] Korsgaard, 2009, p 173.

of agency. What seems a case of endorsement is not actually a case of 'commitment' at all: 'committing' entails being sensitive to the demands of rationality and withstanding rational scrutiny and revisions, which the anorexic's obsession cannot do.

According to this picture then, Korsgaard's model enables us to understand anorexia as a form of compulsion. The kind of view of compulsion that Korsgaard has in mind is one in line with her claim that the tyrannized soul's behaviour is *∂etermined* and it contrasts with the self-determination associated with autonomous agency, which is the result of reflective endorsement and is specifically human. The view that associates anorexia with compulsion enjoys widespread support,[43] and this might seem an advantage for Korsgaard's position. However, I believe that there are substantive problems with this idea. The main issue with this view is that compulsion is understood here as some sort of irresistible urge, which is rooted in neurophysiological and/or psychological processes. This means that, by definition, these agents *cannot* control and choose their behaviour, but rather they are controlled by their obsession. Most importantly, this view clashes with the very fact that, clinically, anorexics do have margins of action, and resistance is possible.[44] It also clashes with the following aspects of their treatment, which would not make sense if truly these patients where *∂etermined* and their behaviour dictated by urges impossible to resist: many people with anorexia nervosa find it hard to acknowledge that they have a problem and are ambivalent about change; treatment tries to make them acknowledge their condition as harmful, win over their reluctance to have treatment and see them effectively engaged in the treatment plan. This outlook and the very possibility of self-management would be ruled out if anorexia were like compulsion, as Korsgaard describes it.

---

[42] Ibid.

[43] See Friederich and Herzog, 2011; and Yaryura-Tobias and Neziroglu, 1983, for correlation between OCD and anorexia. Notably, studies seem to make no clear distinction between OCD and OCDP where the compulsions and obsession (albeit reduced) seem to be ego-syntonic. In general, the understanding of the disorder is complex and complicated by the fact we can find comorbidity not only with OCD but also with other disorders such as depression. See O'Brien and Vincent, 2003.

[44] The view that anorexics lack choice has important consequences on the kinds of treatment these patients can undergo. See Giordano, 2005, pp.33-54 for a discussion of paternalistic interventions in anorexia.

4.2.2 Velleman's model and anorexia

The narrative model struggles to account for the fact that anorexic behaviour is not self-governed because, as we have seen, the model considers that actions are behaviour controlled by one's narrative self-conception. If we hold on to this definition it seems that anorexic behaviour does count as self-governed because anorexics do act in accordance with their self-conception and display narrative coherence. Because of this the model cannot rule out their behaviour as not self-governed in the same way as it does, for example, with cases of weakness of the will. After all, if the coherence with one's self-representation is the condition to count as endorsed, it would seem that anorexic patients indeed are 'endorsing' their values in a way that, for example, Korsgaard's view ruled out. Velleman could exclude that these agents are self-governed in light of the fact that, despite the subjects' treatment of them as reasons, anorexic values fail to amount to actual reasons in two ways: firstly, despite the anorexic's claims and convictions, her beliefs and her behaviour are insensitive to evidence and fall short of standards of rationality that can make them amount to some genuine form of *understanding*. However, this position is not entirely convincing because it is not clear how these standards of understanding relate to narrative coherence and it seems to introduce a criterion for endorsement other than narrative coherence. Secondly, people with anorexia are not really accessing the *real* psychological causes of their behaviour. Under this view it seems that self-understanding does not play a role in producing the relevant behaviour and therefore the behaviour does not result from the right psychological structure. This view commits the model to necessarily associate self-understanding with the real motives of one's behaviour and it leads to a picture of agency that requires standards for self-knowledge that are extremely demanding, as we have seen in the discussion of self-deception in chapter 2. This position is problematic because such high standards of self-knowledge would limit considerably the range of self-governed actions, but also because there seems to be an unclear link in the model between the sense of control and the *fact* of control. In light of these considerations it is not clear how the

narrative model explains the loss of control that is characteristic of anorexic behaviour and because of which anorexics are not self-governed agents.

As we have seen then, both models of reflective endorsement struggle to define exactly the way anorexic behaviour fails to meet the standards for self-governance. In what follows, I will examine a different notion of self-control, developed by Neil Levy and consistent with findings in neuroscience. I will consider whether it is compatible with Korsgaard's and Velleman's model and could be employed to supplement their views on self-control to rescue them from the problems outlined.

4.2.3 Levy's model of Self-Control

In my discussion so far I claimed that neither model is clear on how we can successfully explain the loss of control in cases of anorexia nervosa. In order to better define the notion of control so that it could serve to explain the intuition that anorexics are not self-governed agents, I turn to Levy's work on self-control. The reason for choosing Levy's model is that it develops from concerns about the notion of compulsion that are similar to the ones I raised with Korsgaard. In fact, complaints similar to the ones just raised about associating the loss of control in anorexia to compulsion can be found in the literature on addiction. Despite a standard approach that conceives addiction "as compulsion", several views reject the idea and invoke the same worries I presented above. Levy has developed an alternative, consistent with findings in neuroscience, that takes as a starting point dual-process theories in cognition: his model seems to give a successful explanation of losses of self-control in a wide range of cases, not just addiction (for example, Levy's view targets phenomena like weakness of the will). Because of this, it seems to be a good candidate to explain the difficulties discussed so far. If reflective endorsement theories could integrate Levy's view, this would allow us to modify and better specify the notion of control that Korsgaard and Velleman present. In what follows then I will present Levy's model of self-control, discuss whether it can or cannot apply to anorexia. I will suggest a way in which the model might not seem to apply and what consequences this could have for the views of reflective endorsement I consider here. Then I will suggest a way in which the

model explains anorexia nervosa and turn to the issue of how this model fits with Korsgaard's and Velleman's.

Within the paradigm offered by dual-process theories of cognition, it is thought possible to divide cognitive processes into two basic systems:[45] system 1, which is an evolutionary ancient set of mechanisms whose operations are fast, automatic, effortless, parallel, often modular,[46] operating outside of consciousness and undemanding on cognitive resources; and system 2, evolutionary more recent, distinctively human, comprising mechanisms that are slow, effortful, operate serially (rather than in parallel) and are intentionally or consciously deployed, whose resources are limited and cognitively expensive. Processes associated with system 1 respond automatically to external stimuli and are implicit, reflexive, impulsive, and employ procedures that are contextualized, associative, heuristic; as opposed to system 2 processes, which are controlled, explicit, reflective, deliberative and operate through procedures that are decontextualized, rule-governed, analytic. The two systems deliver different and sometimes conflicting results.[47] Controlled (system 2) processes exert self-regulatory resources that control the effects of automatically activated implicit processes, so that behaviour is kept in line with one's conscious goals.

According to Levy,[48] self-control is a system 2 process and its loss switches us to system 1. This means that a whole set of failures of control should be

[45] What follows is a brief sketch of the dual-process theory as it appears in Levy's work. There is substantial debate around dual-process theories, concerning both the very employment of the terminology of 'systems' (rather than 'processes' or 'types'), and the set of feature that we can list under each one. In this sketched portrait I acknowledge that there are important differences among theories but I consider that I need not to discuss them here. Some will come up during my discussion (for example, the opposition between serial/parallel processes), but for the most part I will try not to complicate the picture I offer here more than necessary. This means that I also will not consider the criticisms directed at dual-process theories: critics point at a variety of unsolved issues which affect dual-processing accounts, from the vagueness of their definitions, to questions about the evidence, to the lack of coherence in the proposed sets of attributes. For a critical overview see Osman, 2004, and for some replies and specifications of the debate see Evans and Stanovich, 2013. Even if we were to successfully associate some dual-process account with theories of reflective endorsement one would need to consider to what extent problems associated with dual-process views can also affect reflective endorsement theories. This is a direction for future work and will not find space here.

[46] Stanovich, 1999.

[47] Evans, 2008.

[48] Levy, 2011.

understood as instances of the broader phenomenon of agents switching from a rational mode of information processing, one that is effortful, slow and draining of cognitive resources, to a more intuitive mode, one that belongs to system 1 processes and that is automatic, less flexible but cognitively under less cognitively load. In making this claim Levy refers to the work on ego-depletion,[49] which is the result of the drawing down of system 2 resources and, Levy claims, explains phenomena like weakness of the will.

> Ego-depletion experiments seem to show that self-control is a limited resource. The psychologists who have developed, tested and refined the ego- depletion hypothesis describe it as the strength model of self-regulation. Self-control is like muscular strength: as we use it, it grows weaker, and can only be restored by rest.[50]

Since system 2 processes are demanding of cognitive resources, it is often the case that one cannot implement them or cannot continue to implement them. When one is under cognitive load or one's attentional resources are depleted due to tiredness, stress and so forth, system 1 dominates and agents act less flexibly, more stereotypically,[51] and their judgments and behaviour will be more strongly influenced by automatic, affect-driven attitudes. So, for example, ego depleted individuals tend to accept weaker arguments,[52] even when counterattitudinal. The theory explains that when some factors (like cravings) come to dominate our attention, keeping our focus requires a cognitive effort that can wear out and deplete our control resources, so that we eventually give in. Levy considers weakness of the will as comprising a judgement shift in conditions of temptation:

> In response to temptation, subjects spontaneously generate or retrieve from memory arguments in favour of weak-willed action. Since they lack the cognitive resources to reject these arguments, they experience judgment-

[49] Baumeister, 2002. See Levy, 2006 and 2011.
[50] Levy, 2006, p. 19.
[51] Kennett and Fine, 2009.
[52] See Levy, 2011, p.142, where Levy discusses the experiments conducted by Wheeler et al., 2007. In this study, subjects were given counter attitudinal arguments. While both depleted and non-depleted subjects were equally convinced by strong arguments, depleted subjects were significantly more convinced by weak arguments.

> shift. They come to judge that the benefits of succumbing to temptation are higher, or the costs of giving in lower, or both, and act accordingly.[53]

Levy develops his view of addiction coherently with these considerations: addicts' autonomy is impaired correspondingly with their depleted resources, in a way that explains, for example, the temporal patterns and likelihood of relapses in substance usage.[54] The idea is that system 1 will be typically biased towards ego-syntonic and gratifying propositions and towards stereotypical, habitual and associative patterns. It will condition system 2 in various ways, for example by directing attention to certain features of the environment. While Levy's work aims at interpreting losses of control in addiction and weakness of the will in terms of a judgement shift, I want to focus here on the wider hypothesis of conceiving self-control as a system 2 process, which is an intuition shared in cognitive psychology[55] and offers an explanation of the psychological mechanisms at work in action production and on the types of information processing that underlie them. We can see how this hypothesis comfortably explains the non-pathological cases like slips and double capture error as lapses of system 2: in these cases the automatic responses take over. Not in all cases in which the agent acts absent-mindedly the switch to system 1 is the result of depletion, but it is important to notice that slips are particularly common in situations in which the subject is distressed in various ways, because he is tired, stressed or emotionally overwhelmed.[56] Nevertheless, as soon as we attempt to associate actions to one system or the other, we find that intentional actions involve an interaction of the two. To understand this, consider habitual and immersed actions: it does not seem that they can be just taken to involve system 1 only, because its processes are non-conscious or sub-

[53] Levy, 2006, p. 143.
[54] Ibid.
[55] Evans, 2008.
[56] It is important to notice however that the role of emotions is not at all clear and would need further explanation, which cannot be given here: some of the examples cited above under "emotional reactions" can be seen as naturally following the pattern just presented. For example, it would take considerable resources and effort to control my speech in occasion of my loved one funeral, and it is not surprising that in this case of intense distress behaviour that is felt both as appropriate for the situation and that corresponds to my automatic response to my circumstances takes over. However, other cases, like the one in which my tensing up results in a poor dive, seem to show how my emotion can cause me to be overly aware of my

personal. These actions are not mere automatic responses to some stimuli, they are things I do intentionally and are not merely controlled by sub-personal mechanisms, as system 1 processes seem to require. Moreover, they are not outside conscious awareness. Driving a car, brushing my teeth, are all activities that we attribute to agents and not to sub-personal mechanisms. The latter are involved in unconscious motor routines, for example, in my adjusting the grasp on a cup I am grabbing,[57] or in tracking unexpectedly moving targets prior to conscious awareness. It is more plausible to hold that intentional actions comprise hierarchies of both conscious and automatic processes.[58] If we take on board this suggestion we can see how these actions are not uncontrolled but also do not involve the level of conscious attention typically associated with system 2: much of the control and regulations of these actions is delegated by the automatic system. These actions can still count as self-controlled because they are counterfactually dependent on the availability of system 2 resources: the monitoring of system 2 processes is not impaired and can inhibit our impulses to act. This makes our seemingly automatic behaviour different from cases of Global Automatism or Anarchic Hand Syndrome: our daily, habitual, automatic actions are sensitive to disruptions and allow system 2 to get a grip on them and resume control in a way that those cases do not allow. The resulting picture of self-control then requires that these attentional resources should be available to me, in order for me to count as being in control. In cases of Global Automatism, where this is not the case, agents do not count as being in control.[59]

I will now consider how this view of self-control explains cases of anorexia nervosa and I will present two tentative possible answers: a first answer urges that this model of self-control cannot be employed to understand anorexia; a second one claims that it can.

---

body and my movements, in a way that disrupts my performance, which would be more optimally conducted automatically and effortlessly.

[57] For studies on the subjects' conscious awareness of these movements see Fourneret and Jeannerod,1998.

[58] Jeannerod, 1997; Pacherie, 2012, pp.100-101.

[59] See Bayne and Levy, 2005.

A first possibility is that Levy's model cannot in fact apply to anorexia. While for some sufferers the model can apply straightforwardly (these are patients that do, at some time, agree that they ought to eat more, and effortfully resist temptation to radical food restriction), other sufferers utterly refuse life-saving treatments. As I stressed before, this is a striking feature of anorexia.[60] These people show no sincere commitment to healthy eating and no shift in judgement seems to occur, even when conditions are life threatening. In denying that Levy's model of self-control is relevant in explaining anorexic behaviour one is faced with two options: one can either hold that these anorexics display the failure of a kind self-control other than the one developed by Levy's theory, or else one can hold that anorexic behaviour should not be understood as displaying a loss of self-control altogether. I will not commit to any of these options but I will discuss how they bear on Korsgaard's and Velleman's models.

The former option holds that there are different kinds of self-control: one correctly identified by Levy and one that anorexic behaviour lacks. This view poses a challenge to reflective endorsement models. Both Korsgaard and Velleman consider that there is *one* property of agents that we must uncover to define autonomous agency: there is one feature of the person's psychology that enables them to draw a single line between the two kinds of agency. If we found that there are different forms of self-control it would be a task for both views to understand how the two relate with each other and contribute to this *one* property defining of full-blown agency. I believe that this kind of enterprise would be difficult because, while Korsgaard and Velleman share with Levy the understanding of self-control as structuring one's psychology, they do not share a fundamental assumption: the difference between their views and Levy's is that he does not equate self-control and self-governance. I will go back to the importance of this association later.

The latter option, one that considers that anorexia is not a disorder of self-control after all, clashes sharply with the intuitions shared by these accounts, for which anorexics are not self-determined and autonomous agents in

[60] Giordano, 2005, p. 235.

important respects. It is important to notice that this is not a problem for Levy because he is not committed to associating self-control with self-governance. Views that, as Korsgaard's and Velleman's, associate the two would therefore be required to question the plausibility of Levy's account and dismiss his notion of self-control; otherwise they could consider how the elements that explain anorexic behaviour can after all be connected to self-control in some way.

A second line of argument is that Levy's model *can* provide an explanation of anorexia nervosa. Anorexia seems to first of all involve a distortion in the agent's body image (or body experience)[61] and we can understand anorexic behaviour as a way of coping with the distress caused by such distortions. It might seem hard to associate anorexic behaviour with the automatic system, because fasting opposes powerful and natural urges to consume food which would naturally be associated with system 1 processes. When agents succumb to temptations towards pleasurable and immediate rewards, we can easily conceive that they effortlessly and automatically switch to such modes of behaviour in cases of depletion and distress; it is much harder, however, to consider how this could be the case for behaviours that seem against one's natural impulses.[62] On the other hand, if we understand anorexics as coping with distress caused by a distortion in their self-image, then we can also understand that resisting the distress caused by one's distorted self-image might be more effortful than succumbing to behaviour that is ego-syntonic and experienced as a rewarding (and pleasurable) way of coping. The involvement of system 1 in anorexic behaviour is apparent if we consider how, in order to achieve their goals of weight loss, anorexics rely on strategies of self-regulation that do not draw on system 2 resources too often. Subjects create the most favourable conditions to bring about their fasting: for example, they avoid eating socially, remove themselves from family and friends and take on routines that maximize weight loss (like counting bites while eating etc.). They

[61] See Bayne and Levy, 2005, for a discussion of a phenomenon that they consider akin to anorexia: Body Dysmorphic Disorder. Where they both appear to be monothematic delusions that are sustained by misperceptions of one's own body. For a general entry on 'body schema' and 'body image' see Gallagher contribution in Bayne, Cleeremans, Wilken, 2009.

[62] Although it is important that fasting and excessive exercise do give a particular type of pleasure, see Giordano, 2005, p. 81

develop new associations and make their fasting into habit.[63] This makes it easier to understand the difficulties patients experience when they are eventually convinced to undergo treatment: treatment needs to target different aspects, from the acceptance of illness to self-management and self-regulation,[64] in order to reduce the sources of distress anorexics' behaviour responds to, and also break habits now performed with a high level of automaticity, which can generate severe draining of the patients' cognitive resources.

This view has important consequences for the approach to anorexia that in the literature is standardly associated with addiction.[65] The reason for this association in the standard view is that addiction is understood as compulsion, and anorexia is supposed to share this compulsive character. Levy's model of self-control aims at resisting views that conceive addiction "as compulsion", when this is conceived as an utterly irresistible desire: the view does not make sense of the very possibility for addicts to attempt and succeed at refraining from the relevant behaviours, nor does it make sense of the clinical practices to treat these disorders. Once we have reconsidered the picture of addiction as compulsion, we can have a different understanding of the parallels between addiction and anorexia. This view of self-control enables us to reconsider the two phenomena as involving the same broader mechanism of switching from system 2 to 1.[66] As a consequence, one can maintain the association between addiction and anorexia without assuming that they entail compulsion, which was the ground of their standard association.

---

[63] "The great thing, then, in all education, is to *make our nervous system our ally instead of our enemy... For this we must make automatic and habitual, as early as possible, as many useful actions as we can,* and guard against the growing into ways that are likely to be disadvantageous to us, as we should guard against the plague" (James, 1890, vol. 1, p.122, italics in original).

[64] For a review of the possible adjustments of coping strategies in chronic illness, including anorexia, see de Ridder *et al.*, 2008.

[65] Giordano, 2005, pp. 76-80.

[66] There are considerable similarities in the ways to effectively treat these conditions that offer patients new ways of copying and strive to forge new associations by engaging deliberately and repetitively in alternative behaviour and most importantly relying on peer support, family therapy, forming interpersonal attachments and giving new prospects on 'the good life'. See Pickard and Pearce, 2013, for a good overview of the strategies adopted with addicts, once we have abandoned the view that addiction is compulsive.

Can the notion of self-control as a system 2 process be integrated with the models of reflective endorsement developed by Korsgaard or Velleman? There would seem to be a prima facie difficulty here, since Levy's theory does not try to define agency as such, but rather seems to focus on the mechanisms of information processing and their relation to the executive functions. However, this could be considered just a superficial concern: the fact that the theory is not in itself concerned with constitutive conditions of human agency does not mean that it cannot be consistent with views that seek to establish those conditions. One could argue that we can integrate reflective endorsement within this picture offered by cognitive science: this requires us to evaluate how the constitutive standards of agency specify the kind of system 2 process that self-control requires. Different conceptions of reflective endorsement, such as that based on the categorical imperative or on narrative coherence, would then define the rules and hence the normative, not merely associative, relations reflective endorsement establishes.

Moreover, by relying on dual-process theories of cognition, this view of self-control seems to map the division between kinds of agency we have discussed so far: firstly, system 2 is distinctively human, as opposed to system 1 with which we can credit other mammals; secondly, system 2 allows for the kind of activity that reflective endorsement theories try to define, as opposed to the behaviour that results from triggering by environmental stimuli that characterizes system 1. Moreover, the idea of self-control being a system 2 process distinguishes it from the kind of motor control that we find in system 1: this is important because it is not the impairment of motor functions that concerns twofold theories of agency in explaining the conditions for self-governed behaviour. The performance of mere activities, for example, my instinctively grasping an object thrown at me, requires that I have motor control of my movement but this kind of control is not what is necessary for the movement to count as a full action for these theories. Lastly, Levy defines the conscious information processed by system 2 in terms that fit with the requirements of both models:

> Information is conscious in the right kind of way, I suggest, when it is *personally available*. But what is personal availability? Information is personally available when the agent is able to effortlessly retrieve it for use in reasoning *and* it is occurently online, actually guiding behavior or mental processes, prior to retrieval. Both conditions are needed. If the information so retrieved only guides behavior as a consequence of retrieval, it does not count as personally available *until* it is retrieved. But being online is insufficient for personal availability because unconscious states often guide behavior without our being aware of their contents or their effects. It is the conjunction of effortless retrievability and being online that is needed. It is likely that online information is available for effortless retrieval when it is available to a broad set of systems, including systems involved in reasoning; availability for retrieval is a reliable indicator of availability for use in reasoning.[67]

My discussion of INS and INS* in the narrative model pointed at the fact that in actions one's narrative understanding needs to be both readily accessible and guiding our actual behaviour. The ideas seem to be akin to those of effortless retrieval and being online presented in the passage above.

It seems prima facie possible, then, that Levy's view of self-control is consistent with these models of reflective endorsement. However, as we shall now see, Korsgaard's view is utterly incompatible with a purely procedural view of self-control like Levy's, while Velleman's view seems more promising in being associated with it.

First I will discuss the way Korsgaard's view contrasts with Levy's. Levy's notion of self-control is in line with Baumeister's considerations that see willpower as akin to muscle strength. The view of self-control as a system 2 cognitive processing capacity offers a purely procedural notion of self-control that does not capture the normative dimension required by Korsgaard. Recall our discussion of what seemed to go awry in cases of anorexic behaviour according to her view: the person's behaviour is indeed controlled by her self-representation, a representation that conceives herself as fat and lightness as a value. However, according to Korsgaard, this self-representation just is not the *right* one to allow true guidance, because it cannot be the expression of the categorical imperative. This is why, for Korsgaard, the impression of being in control and having the right phenomenology of agency is just illusory. In her

[67] Levy, 2013a, p. 214.

model, other agents that fail to be self-controlled also have the wrong self-representation, one that clashes with the categorical imperative: the selfish person sees himself as an exception to moral concerns; the wanton does not differentiate himself from his desires, thus not adopting a practical identity whose principles can fit with the categorical imperative. Korsgaard's proposal aims at grounding the conditions for *moral* agency on the conditions of agency *itself*: in her model self-control is the condition to achieve unity of agency and the protagonists of all these examples fail to attain this unity because their principles clash with the one principle constitutive of agency, the categorical imperative. This means, according to Korsgaard, that the unity of purpose of their lives is also defective. The procedural view of self-control articulated by Levy does not fix and cannot possibly respond to the substantial standards of rationality and unity required by Korsgaard. Attempts to impose this normative character on the procedural notion of self-control derived from Levy's account seem entirely arbitrary and ad hoc, since nothing in the theory seems to support these requirements. Even if it turned out that self-control and system 2 processes play a special role in defining the conditions of those actions that, for example, bear a relation to moral responsibility,[68] the kind of normative results the theory could enjoy would not have the constitutive relationship with the conditions for agency required by Korsgaard. According to Korsgaard, normativity is grounded in the nature of agency because the categorical imperative is the constitutive principle of both agency and morality. This means that the conditions for self-governed actions are also the conditions for good actions, in Korsgaard's Constutivism. While normative considerations can be drawn directly from the conditions for self-governed and self-controlled behaviour in Korsgaard, they cannot be drawn directly from Levy's notion of self-control. From Korsgaard's perspective, it is not just the case that Levy's notion of control fails to account for cases that require a normative dimension of self-control but might be correctly applied to others:[69] Korsgaard's view aims at finding that *one* property that defines autonomous agency and it requires a univocal notion of self-control. A purely procedural notion relying

[68] This seems to be Levy's view, see Levy, 2013a, Bayne and Levy, 2004.
[69] As in Kennett, 2013.

on different types of information processes just cannot be of any relevance for Korsgaard.

Velleman's position seems easier to associate to this notion of self-control: the standard of narrative coherence does not immediately bring with it normative considerations as the appeal to the categorical imperative does. Moreover, despite the fact that, at this stage, the association between narrative coherence and system 2 processes of self-control is a mere stipulation, there is considerable psychological evidence that stresses the role of consistency and predictability,[70] and this seems to go well with the role assigned by Velleman's model to the agent's cognitive drives. There is no reason to think that this form of reasoning has to have any narrative element, but this is a concern that at present can be put aside. Levy thinks that consistency links propositions semantically and does not need to depend on agents valuing and pursuing consistency per se. This is a position similar to Velleman's, who stresses how cognitive drives of self-understanding and consistency regulate reasoning automatically, while remaining for the most part in the background both of action and thought.

Despite these initial points of agreement, a closer look to the details of both views suggests that Levy's account of self-control also may not suffice to address problems we found in Velleman because Levy seems to equate system 2 processes with processes of deliberation and makes the following claim:

> Deliberation requires the rule-based processing that is the domain of consciousness; the associative processes generated by unconscious attitudes should not be thought of as reasoning. Deliberation, moreover, is *deliberate*: it is an exercise of agency. But the exercise of epistemic agency – the assessment of reasons – is the domain of consciousness. We can only assess – for consistency and for plausibility – what happens on stage (and offstage relatively little testing for consistency occurs).[71]

[70] For a review of the evidence supporting the idea that conscious attitudes are automatically integrated and updates in ways that reflect relations of consistency see Baumeister and Masicampo, 2010; and Levy, 2013a.
[71] Levy, 2013a, p. 222.

A first striking point that emerges is that Levy considers deliberation as an action in its own right, one that is effortful and is undertaken intentionally. I have already discussed this idea when presenting Korsgaard's model and there is an interesting consideration to be made on this point: since Levy is not concerned with defining the conditions of agency through these claims about system 2 processes, the issues I raised earlier about identifying reflection with deliberation need not arise for his model.[72] As we have seen, Velleman attempted to define a further dimension of reflection by distinguishing deliberative and perceptual reflection. The reason behind this distinction was both to allow reasoning to function, so to say, passively, automatically, and also to account for the most daily, habitual and immersed actions. The two functional roles that were necessary for the self as narrator to play were the one of the deliberator and the one of the supervisor. I raised doubts about the possibility for the narrator to play both this active and passive role. Levy here seems to equate the processing of system 2 with an intervening role, rather than with a merely monitoring one. Levy thinks that automatic actions of the sort that Velleman explains through perceptual reflection "are (by definition) not directly monitored, …they are indirectly monitored to the extent that the agent remains ready to move to conscious forms of agency should that be necessary".[73] The definition of 'monitoring' is a debated and delicate issue for dual-process theories of cognition in general, but, for the present discussion, it is sufficient to consider the reasons why Levy's system 2 processes cannot be credited with Velleman's passive supervisory role.

This becomes clear if we understand the sense in which Levy is committed to the idea that system 2 processes operate sequentially or serially, rather than in parallel. As noted earlier, dual process theories generally hold that the analytic system monitors the output of the heuristic system (though according to Levy they do so indirectly). When a conflict with analytic knowledge is detected, the

[72] These were, mainly, the regress that results from this identification and the impossibility to consider behaviour which does not involve deliberation as action. Moreover, notice that it is possible for this view to consider that the phenomenology distinctive of deliberation results from the effort and the experience of the activity of deliberating itself. This is not an idea that reflective endorsement views of agency can support: they consider an actual difference in the *kind* of agency we can attribute to reflectively endorsed behaviour and this difference cannot be reduced to just an augmented sense of effort or activity.

analytic system will attempt to intervene and inhibit the prepotent heuristic response. Levy, along with Kahneman and Frederick,[74] and Evans[75], conceives a serial characterisation of the interaction between the two systems: one initially starts by relying on the heuristic system and the analytic system only intervenes at a later stage. In contrast, Sloman[76] holds that both routes are supposed to be simultaneously computing a solution to a problem from the start. Velleman's driver/passenger seat metaphor aimed at giving the sense that the whole ride can be seen as the result of one's practical reasoning, to the same extent that some result can be credited to a supervisor, even in cases where no intervention was needed. Because of this, I believe that it would seem more apt to associate Velleman's view with one, like Sloman's, that favours parallel over serial processing.[77] It is important to notice, however, that a parallel processing architecture seems to violate the principles of cognitive economy: parallel models offer a picture in which the analytic route is always already engaged right from the start, and we have seen that this route is both slow and demanding on cognitive resources.[78]

Moreover, there are two further aspects in which Levy's and Velleman's models differ. Both concern the notion of conscious access the two views assume. A first aspect, which I won't expand in detail, is the fact Velleman's model seems to be more demanding than Levy's on the self-knowledge required.[79] The second aspect concerns the understanding of access and monitoring through which analytic, system 2 processes can get a grip on behaviour led by system 1. System 2 processes can effectively stop the route of behaviour and most importantly, they can do so for reasons, accessible to the

---

[73] Bayne and Levy, 2004, p. 214.
[74] Kahneman and Frederick, 2002.
[75] Evans, 1984.
[76] Sloman,1996.
[77] While considering other dual-process accounts in relation to Velleman's model can be an interesting way forward, it is not a subject that will be pursued here.
[78] For an overview of the discussion about monitoring and the alternatives to this serial/parallel dichotomy see De Neys and Glumicic, 2008.
[79] I already presented Velleman's implausibly high standards of self-knowledge in chapter 2. Levy rejects unrealistically demanding standards that I have attributed to Velleman in the previous chapters. This is congruent to the fact that the notion of access I presented here is developed by Levy within him attempt of considering the relationship between consciousness and moral responsibility. As we have seen, this access is, for Velleman, connected with the

agent. This view can lead, in Velleman's account, to worries concerning action individuation that do not arise in Levy's. If actions are self-controlled, and self-control requires system 2 processes, it seems that only at the stage in which system 2 intervenes can the same act, even a same movement, count as action because only then does it fulfil the conditions for agency. Since the kind of intervention we can here credit to system 2 is necessary for agency on Velleman's view, it is unclear that this movement could count as action all along and what it means for some mere activity to *become*[80] an action. This is not a worry for Levy's view because he does not establish a constitutive relation between conscious control and agency, and hence between the access required for reasoning and agency. By contrast, Velleman's project concerns the conditions of agency and it seeks to better Davidson's causal model by explaining how that which causes the action constitutes the agent's reason for it. In Levy, access is necessary for system 2 processes to get a grip on behaviour but there is no need to identify reasons with causes with the aim of fulfilling constitutive conditions for agency. Because of this, it is possible for Levy to hold that someone's behaviour goes from being uncontrolled to be controlled, but this has no bearing on the status of such behaviour as action: it does not change the fact that one is acting.

In this section I tried to show, through the discussion of anorexia nervosa, what self-control really amounts to in Velleman and Korsgaard. I showed how their conception of self-control is not satisfactory in its explanation of the way in which these agents are not self-controlled and self-governed. I tried to consider whether one could integrate them with Levy's influential account of self-control which argues that self-control is a system 2 process and that losses of control show a switch from system 2 to system 1 type of processes. I considered that Korsgaard's view requires more than the purely procedural notion of control theorized by Levy. I also argued that Velleman's view promises at least the possibility to fit with accounts of control like Levy's, which take control to be a system 2 process, but it also requires a way to

very possibility of agency and it involves one's narrative self-representation. See Levy, 2013a, pp. 225-226.

characterize monitoring that is different from Levy's. I do not claim that this analysis is in any way conclusive or leads to some knockdown argument against the notions of control deployed by Korsgaard and Velleman. However, I hope I have highlighted the substantial unclarity of this notion in the two views and the difficulties in matching it with one employed in empirical work. I have pointed at how reflective endorsement theories present difficulties in explaining anorexic behaviour because they associate the notion of self-control, which requires agents to present a certain psychological structure, with self-governance. In my final section, I will explain further why equating self-control and self-governance is in itself problematic.

4.3. Self-governance and Causal Explanation

The above discussion showed in what way I believe the theories examined hold an obscure and problematic notion of control. In this section I want to present a consequence of this fact.

The obscurity of the notion of control bears on the overall project of defining agency because self-control is identified with self-governance, which is what characterizes the specificity of human agency, and the result is that self-governance also remains an unclear notion. In absence of further clarification, I believe that this identification of self-governance and self-control is problematic. I will claim that, as a result, we have reason to question even the very definition with which I started the chapter and according to which an action being self-governed really can correspond to it being guided or controlled by either one's narrative self-representation or by a principle with which one identifies.

To see this we can take as a starting point the reason why Velleman claims that Davidson's account failed to give us a depiction of the agent's role in action. Davidson's account, Velleman urges, is not an account of actions, but of mere

[80] Velleman employs this jargon in 2000, p.191.

activities: it does not capture what "distinguishes human action from other animal behavior".[81] For Velleman, since Davidson's view of agency has no place for self-governance, it does not explain the way in which the agent participates in the action. The causal story makes us lose sight of the agent, of why he did what he did. We have seen how Velleman tries to compensate for this deficiency by introducing psychological elements that, while embedded in the event causal order, can be fundamentally linked to one's *self*, so that the behaviour brought about counts as *self*-controlled. As we have seen, his solution was to add some psychological element, congruent with the causal account, that could "play the role of the agent". In this way Velleman's conception of action as behaviour controlled by one's self-representation fits within the general strategy of the standard causal theory of action: it maintains that giving an explanation of an action in terms of how it has been caused by some mental state that plays the role of the agent just is equivalent to explaining actions as brought about by self-governed agents. What is not clear though is how the proposal of adding an extra state to the standard story can solve deficiency of Davidson's view to account for self-governance. How can some state of the agent reinstate the agent in the picture? It is unclear how Velleman's theory succeeds in explaining why the agent did what he did, because it still gives us a story about how some psychological elements cause events that count as actions.[82] In previous sections I have claimed that the notion of control is obscure, and as I said a consequence of this is that the notion of self-governance is also unclear. As a result, instances in which the two seem to come apart remain unexplained and the strategy of adding certain

[81] 2000, p. 124.

[82] This worry is related to Jennifer Hornsby's claim that causal standard theories of action leave the agent out of their picture of agency by seeing the issue of action as one of defining some specific causes in operation. According to Hornsby, theories that recognize the shortcomings of the standard theory, but try to solve them by adding new mental states, also fail to understand what truly goes amiss in the standard story and they do not succeed in accounting for agency. See Hornsby, 2004, p. 2. These theories devote their analysis to identifying the psychological causes for certain events and assume "that citing states and events that cause a bodily movement carries the explanatory force that might have been carried by mentioning the agent" (ibid., p. 13). What results from this position is a model that does not convey the role the agent played in action: the difference that is never explained is the one between someone bringing something about and some psychological state causing some event. While I present here the criticisms that echo Hornsby's worries about the standard theory, I do not mean to commit to the positive thesis she presents. In my conclusion it will become clearer what lesson I think we can draw from the considerations I raise here and this

mental states to the standard story does not help to explain them, but, in fact, it only raises further doubts about this identification.

We can see how equating self-governance with this kind of control is problematic by examining the way in which I have claimed both Velleman's and Korsgaard's models limit the range of actions.[83] The problem with excluding cases of mundane or ordinary actions (as Korsgaard's model does) or acts resulting from self-deception (as Velleman's does) is that we have the intuition that these kind of acts should count as actions: we do not see agents performing mundane actions as not self-governed, nor do we think that the self-deceiving agent carefully planning to lie to his parents is acting in an almost animal-like manner that sees no participation on part of the agent. Nevertheless the models exclude these kinds of behaviour because they do not involve the right psychological elements and do not have the right causal history: because of this, these agents' actions do not count as controlled by their self-representation in the ways in which the two models require. In light of this I believe that we can question how the strategy of picking out a certain state to play a certain causal role can be of aid in reintroducing the agent in this picture of agency and explain self-governance.

Moreover it is important to see that this kind of account is *bound* to lead to the implausible restrictions on the range of agency I have highlighted, because it holds that there is one clear line to be drawn between merely purposive behaviour and full-blown agency. This means that according to these views there must be one kind of state that marks actions as such. The limitations on the range of actions, then, spring from restricting the range of the mental states that could be candidates to play the agent's role. If we consider that being self-governed agents just cannot amount to our actions being controlled by some kind of mental state, it becomes clearer why these proposals fail to solve the original problem afflicting the causal theory of action. Moreover, the fact that

---

will lead me to consider some desiderata for a theory of agency that can avoid the difficulties I attributed to reflective endorsement theories.

[83] I have introduced the present point in relation to Velleman, but, as I will show shortly, I believe that it also applies to Korsgaard's view and it is because of this that I bring her model here as an example next to Velleman's.

these models define a very limited range of agency is problematic because it is implausible that the defining property of human agency derives from a very limited range of occasions in which behaviour is produced through the right psychological structure.[84] I believe that this problem also mirrors the one I presented in the conclusion of the last chapter, since, for Korsgaard and Velleman, self-governance is essentially linked to the agent's self-consciousness. As we saw there, Korsgaard and Velleman define self-consciousness in terms that require a richness and a complexity that we do not really need to capture the specificity of human self-consciousness. The upshot is that these views not only impose limitations on the kind of behaviour for which this high-level and demanding form of self-consciousness plays a role, but they also impose a stance that does not capture the practical and essentially first-personal character of agency.

It is important to notice that the association between self-control and self-governance is at the heart of these views' attempt to connect the philosophy of mind and action to the discourse around self-governance, which is thought fundamental to connect issues in moral psychology to the practical capacities we can ascribe to human beings (for example in order to connect the specifics of human agency to issues of ethical normativity). Views that do not share this aim, even if they were to define self-controlled action along similar lines, that is, as action controlled by some particular kind of state, need not be afflicted by the present worry. Such views could, for example, distinguish self-controlled actions and automatic actions because they have a different psychological and causal history, but still hold that both kinds count as actions and they can fail to be controlled, without their agent being less than a self-determined agent.

Let me explain the extent to which I believe that Korsgaard's, as well as Velleman's view, is affected by this issue, namely, that the identification between self-control and self-governance is doubtful and that adding mental states to the causal story does not provide an elucidation of why someone

[84] Ibid., p.10.

whose behaviour is controlled by his self-representation counts as a self-governed agent. It may be evident how this point applies to Velleman's model, since he explicitly calls for certain states to "play the role of the agent". But it might seem that the problem does not affect Korsgaard's view: after all, Korsgaard's stress on the active part played by the agent in her own action is intended to show how action must be seen "as an expression of my self as a whole, rather than as a product of some force that is at work on me or inside me".[85] In such claims, Korsgaard might seem to better answer to the worry just raised, rather than being its target. Nevertheless, I believe that Korsgaard's position is affected by the criticism I just discussed.

Korsgaard explicitly suggests the consistency between her views and those which, like Velleman's, try to supplement the causal theory of action.[86] This in itself is not particularly important because one could claim that her normative constitutional view can avoid the problems I just described, since its consistency with causal theories is not at all necessary for its formulation. However, I think that the reasons for which Korsgaard's model is consistent with causal theories run deep and are grounded on assumptions that ultimately expose it to the present worry, even if we do not assimilate it directly with causal accounts.

Korsgaard claims that the view that actions are movements caused by the subject's mental states and representations turns the story of agency into one that sees some chain of causes "running through"[87] the agent, and thus fails to account for the specific activity of the agent. But she also thinks that these shortcomings of causal theories can be amended when we consider actions as movements that have their source in the *self*, in one's essential identity. Korsgaard, then, conceives actions as caused by mental states that count as expressions of the agent's essential identity and holds that, therefore, these actions can then be considered as caused by the agent. This means that, for Korsgaard, actions do not just result from mental activity, but from a mental activity that is normatively constituted. The view still assumes the standard

[85] Korsgaard, 2009, p. 18.
[86] For Korsgaard's discussion of the approaches to agency that she calls "natural" and the relations with her theory see Korsgaard, Forthcoming.

story of agency, but what it requires to cause actions are not mere beliefs or desires: they are principles,[88] whose workings can still be conceived as mental operations, while they can also account for the normative and active character of full-blown human action. This is so because they are bounded by one's essential nature as a rational agent. But, although Korsgaard claims that this introduces a different kind of causality,[89] this different kind of causality does not reinsert the agent into the picture. In fact, it is not clear to me that it amounts to a different kind of causality at all, but rather it is a matter of some event having some metaphysical property (autonomy), which it gains from the holding of a certain psychological relation: one of unity between one's motives and principles. Korsgaard's stress on the role of the *whole* agent is meant to rule out the possibility of heaps of unrelated states resulting in behaviour that can count as action. Nevertheless she still assumes some principles that are elements of the agent's psychology and provide the unity necessary, and are states and events that cause actions.

4.4. Conclusions

To sum up this chapter, I hope to have shown that Korsgaard and Velleman do not present a particularly useful way to think about the notion of control: Velleman's model failed to explain the relation between control and agentive awareness, while Korsgaard's conceived it in very restrictive and demanding terms. They both also struggle to provide a solid account of the psychology of self-controlled action and Korsgaard's model, in particular, seems to require something other than the purely procedural notion of control we can find in cognitive psychology. Finally, I have argued in the last section that defining action as behaviour controlled by one's self-representation does not explain what it is for agents to be self-governed and self-determined. This view of self-governance seems to consider that it is necessary to equate agents with some

[87] Ibid., p.10
[88] A principle seems to take the form of deciding whether an incentive bids, and, in this view, a decision is still a mental event. It is important to note that in light of these considerations, issues of regress that I discussed in chapter 1 seem inevitable.
[89] Korsgaard, 2009, p. 75.

mental states (either one's narrative self-representation or the principles defining one's essential identity), because otherwise agents will seem as some distinct, mysterious and ineffable element in the metaphysics of agency. I do not believe that this attempt truly succeeds in explaining the connection between control and self-governance and, moreover, I do not think we need to endorse either of these positions. In my conclusions I will hint at a possible alternative.

Conclusion

I started this work from the intuition that human beings are agents, and that they are agents in a rather different way than animals are.

I presented an influential way to think about what makes human agency different: the idea that human beings are capable of a different kind of agency and that self-governance is the differentia that characterizes it. Self-governance enables human agents to have special authority over their behaviour. Reflective endorsement theories ground the conditions for the different authority we can enjoy over our behaviour in our reflective capacities. This idea, in turn, leads these views to attribute a key role to self-consciousness.

I have discussed the work of two prominent philosophers who operate within this framework: Christine Korsgaard and David Velleman. My analysis took into consideration their description of actions, as different from the mere purposive activities which we can also credit some other animals with. I showed that these models are not satisfactory and the account that these authors give of human agency is flawed. In chapters 1 and 2, I argued that they restrict unduly the range of action, and that this is due to their understanding of what reflection amounts to. In Korsgaard's model, action seems to necessarily require deliberation and this excludes a substantial range of acts which are not deliberated, but which we would intuitively think of as actions. I also argued that we cannot understand Korsgaard as employing the term deliberation just as a synonym for 'practical reasoning' because her model

puts an important weight on the agent's experience in a way that this identification would exclude. Because the agent's experience is important, the model needs to provide a convincing explanation of it, but Korsgaard is ambiguous in her definition of the notion of 'identification'. On the one hand, she seems to see it as a feature of the agent's experience and, on the other hand, it seems to refer to the psychological process that determines the structure of the agent's will.

As for Velleman's model, I claimed that the appeal to the role of narratives and the attempt to describe reflection not merely in deliberative terms allows his model to account for a wider range of actions than Korsgaard's. For example, Velleman seeks to include within the range of actions our mundane and ordinary acts, which were excluded by Korsgaard's account. I claimed that this attempt is partially successful but still ultimately inadequate: this is because, on the one hand, there is a tension between the supervisory and narrative role he credits the self with, and, on the other hand, because his overly high requirements for self-knowledge exclude self-deceptive behaviour from the range of actions. As a result of this, the range of behaviour that counts as action is, as in Korsgaard's case, seriously limited. In short, both theories lack of explanatory power.

In chapter 3 I discussed in more detail the notion of the narrative self, understood as a narrative self-representation that the agent *has*. I showed that Velleman's appeal to narratives is problematic because these are always interpretive in character and do not secure the link between one's understanding and the actual causes of one behaviour that Velleman's model requires for action. It is a limitation of Velleman's theory that it does not admit any slack between the two, while the appeal to narratives seems to naturally allow for misinterpretations and biases.

I also argued that the fact that both Velleman and Korsgaard understand the self-consciousness required for agency in terms of providing an objective self-representation is problematic. I argued that this understanding rests on an overly intellectualistic and limited conception of human self-consciousness and

prevents both models from providing a realistic picture of the actual specificity of human self-consciousness.

In chapter 4 I examined the notion of control required for full agency by both views. I claimed that neither view is satisfactory in explaining the connection between one's control and agentive awareness. Velleman's account is ambiguous, and ultimately fails to explain the relation between control and the experience of agency. Korsgaard's account is very restrictive. I discussed also the claim that self-controlled action requires the right psychological structure and I tried to show that the models fail to account for agents who have the right psychological structure but cannot be considered self-governed. As a result, the notion of self-control necessary for the agent to have the right psychological structure remains rather obscure. Finally, I argued that both models misrepresent the role of the agent in action: the very definition of action as behaviour controlled by one's self-representation misses out on what it is for agents to be self-governed and self-determined. This view of self-governance seems require equating agents with certain mental states that cause the action.

In light of these considerations, I can now go back to my original question, which was whether it makes sense to distinguish two kinds of agency, one of which expresses the agent's self-governance and is specifically human. This thesis offers the following answer: in the influential variants of reflective endorsement elaborated by Korsgaard and Velleman, it does not. To this extent the results of this work are primarily critical and might seem quite modest. This is apparent if we consider the overall picture I presented in my introduction: reflective endorsement models belong to an approach in philosophy of action that strives to find and isolate that one feature that makes some activity into an action. There are various commitments that these views have: first, they rest on the idea that actions are a species of activity, and secondly, they hold that there is *one* feature that gives us the differentia in terms of which we can understand the contrast; this differentia is self-governance. Moreover, they interpret self-governance in terms of reflective endorsement. Velleman and Korsgaard build their models from their

understanding of reflective endorsement. It is only this very last step that the thesis has assessed: the differentia as conceived by Velleman and Korsgaard is not satisfactory. This conclusion does nothing to debunk the other assumptions I just presented and thus it gives no ground to dismiss reflective endorsement approaches in general, nor to question this overall picture of agency.

Nevertheless, I believe that the discussion thus far opens the way to two further directions of analysis: firstly, the shortcomings I have highlighted thus far can be used as guidelines to test other reflective endorsement theories. These shortcomings rest on fundamental assumptions concerning the notion of control and self-consciousness. These assumptions result in systematic difficulties concerning a possible account of agentive awareness, of self-control and the narrow restriction of agency the models are committed to, which, for example, results in an unsatisfactory account of ordinary, mundane actions. These considerations provide us with some guidelines to assess other reflective endorsement theories. We can check whether these assumptions apply and, if so, we can expect similar difficulties to arise, namely, ambiguities in handling agentive awareness or undue restrictions of the range of genuine actions. Alternatively, we can examine views presenting similar problems[1] and uncover their connection with the same flawed assumptions I highlighted in this work.

A second, and more positive, result of this discussion is that we can distil from these considerations the desiderata for a plausible theory of agency, as follows:

a) Specificity of Human Agency. The idea that human actions are in some way different from animal actions should be preserved and a theory of agency should be able to explain such a difference.

b) A plausible range of actions. While we sometimes act on the basis of a deliberated decision, a lot of our mundane behaviour does not involve deliberation, nor does it express our deep values. A theory of agency should do

[1] For example, Bratman's. See Hornsby (2004) for a discussion of a theory that also narrows substantially the range of actions.

justice to the idea that mundane actions are not defective and *are* genuine actions.

c) Clear conditions for agency. A theory of agency needs to be able to establish whether some instance of behaviour is or is not an action.[2] I have argued that Velleman's theory, for example, falls short of providing such criteria: this emerged from my discussion of self-deceiving agents in chapter 2, but also from the analysis of the notion of conscious control in chapter 4. In comparison, take for example Burge's definition of action as "coordinated behaviour by the whole organism issuing from the individual's central capacities, not purely as sub-systems"[3]: such a definition delivers a firm distinction between actions and things that happen to the individual, or processes that occur within the individual.[4]

d) A realistic conception of self-consciousness. This is something that I have argued the models under consideration fail to provide. Both models emphasized the role of one's objective self-conception, conceived in terms of one's narrative or in terms of practical identities, with the result that these views flatten one's purely first personal perspective to the one we can associate with other animals capable of consciousness. This seems to give a picture in which the capacity for self-consciousness simply adds on this first personal perspective. This view is not realistic, because it fails to take account of the fact that our rational and self-conscious capacities are not simply added on, but they thoroughly modify, that first-personal perspective altogether. The first-personal perspective is already specific to human self-consciousness and its understanding needs further development. The demand for a realistic account of self-consciousness brings with it that one needs to go beyond the generic claim that it is self-consciousness that makes the difference in human agency.

---

[2] This of course allows for error, since even if the criteria are clear, it is possible to mischaracterize a single case.

[3] Burge (2009), p. 260

[4] I am not here claiming that Burge's definition here manages to account for what is distinctive and specific of human action. In fact, this definition of agency as it is stated here, does not fulfil *a)* and is developed by Burge to account for the primitive agency we find also in other animals. Nevertheless, what I wish to emphasise here is rather that despite the fact that this definition of agency might fail, as it is, to account for other desiderata we include in a theory of action, it does deliver clear results and establishes clear conditions for agency.

In order to do this, a theory should not overlook those less than full-fledged dimensions of self-consciousness and avoid confining self-consciousness to a very high level of cognitive sophistication. If, as reflective endorsement theories hold, self-consciousness is the differentia that allows us to define what it is to be a human agent, we need an account of it that explains plausibly the developmental evolution in the cognitive abilities specific to humans and, in ontogeny, from the cognitive skills and abilities of normal human infants to the relevant forms of higher capacities we credit adults. In fact, if a given cognitive capacity is psychologically real it must be possible to explain how an individual in the normal course of development can acquire it. While reflective endorsement theories, and philosophical theories of action in general, might not be interested in providing an account of how the notion of self-consciousness emerges, both in ontogeny and in phylogeny, one's conception of self-consciousness must at least be coherent with such an account. Moreover, a realistic account of self-consciousness should not set implausibly high standards for self-understanding and should allow for possibility of the opacity of self-knowledge.

e) An integrative account of agentive awareness. Agentive awareness is complex and, as we have seen, it has different dimensions.[5] With Bayne, I believe that "the experience of acting is not exhausted by the conscious judgments one might have about what or how one is acting, if indeed one has any conscious judgments of that nature at all. Acting may frequently involve conscious judgment about what one is doing, but agentive self-awareness is not primarily a matter of judgment"[6]. By understanding the agentive awareness necessary for full-blown agency in terms of a limited form of SA2, Korsgaard and, more ambiguously, Velleman, seem to conceive it in terms of judgements. I have argued that their conception of agentive awareness is not satisfactory. A plausible theory of agency needs to explain the relation between one's experience and judgements. This is something that, because of the clear cut

[5] This issue is of course connected to an adequate understanding of self-consciousness. If we do not conceive of the importance of this first-personal perspective for self-consciousness we cannot make sense of the complex and fragile phenomena that self-awareness is (for example in relation to first-personal reference) and hence cannot explain the role, albeit crucial, of self-consciousness in human agency.

that these models draw between kinds of agency, their theories actually cannot offer. In order to provide a full account of agentive awareness, one needs to provide an explanation of their interaction. In limiting the agentive awareness relevant for agency to SA2, one loses sight of a whole level of the experience of agency, (SA1), and therefore cannot explain the relation between these two levels. A theory that avoids this approach can, instead, understand the role of the experiences that usually ground one's judgements. And it also opens to the possibility of exploring the judgements' bearing on one's experiences, for example through some form of cognitive penetration.[7]

In the attempt to do justice to the first of these desiderata, the reflective endorsement theories I have examined failed to meet others, in a way that, for example, a more permissive conception of agency would not. In what follows, then, I will make some sketchily suggestions about how such a theory could be developed.

By a permissive theory of agency I mean one for which agency is a capacity of organisms to bring things about and which has different characteristics depending on the other capacities the organism is endowed with and which define it. In light of this, a permissive theory of agency need not have a special notion of agency for humans: it simply holds that human agents are human beings who act. In this sense there need not be any different *kind* of agency, enjoyed by human beings only. Such a theory does not deny that human actions are different from those of other animals, or that human actions present characteristic or specific features, but, in allowing for such a difference, it does not claim that there is a different kind of agency, characterized by a fundamental constitutive property (such as autonomy, according to Velleman's and Korsgaard's Kantian projects). In fact one could argue that actions can be seen as exhibiting different characteristics and credentials according to the different capacities we attribute to their agents. While this seems intuitive and not particularly remarkable, it does allow us to understand differences in the

[6] Bayne (2011), p. 359.

actions performed by agents displaying different cognitive capacities in a continuum, according to a principle of gradualism, rather than with a precise leap where the faculty of *reason* comes into play, as it happens in projects inspired by Kantian faculty psychology. Moreover, such a permissive theory can still allow self-consciousness to be a distinctive human capacity, without considering that it provides a clear-cut threshold to define kinds of agency. The idea that human agents are human beings that act has the consequence that what they do is specific of their form of life throughout: it means that even those doings which the models I examined considered as mere activities, rather than actions, were expressions of human agency, nevertheless. If we understand that the specificity of human actions is brought by the employment of the distinctive cognitive capacities for self-consciousness, and we see that in every, even minimal, instance of human agency there is a relevant dimension of this self-consciousness, even when it just involves a first personal perspective[8] rather than an objective self-conception, than it seems that really every minimal instance of agency is specifically human. One upshot of this view is the possibility of a more integrative understanding of agentive awareness. Talking of agency in a unified manner (without relying on a distinction between *kinds* of agency) allows for a unified conception of agentive awareness, which can comprise different aspects, both SA1 and SA2.

The full development of such a permissive account of agency is not the aim of this work, and would require a more detailed assessment of the desiderata I outlined, which is not something that can be accomplished here. Nevertheless I do wish to briefly draw some considerations on important upshots of this kind of conception of agency. Firstly, the kind of ownership of behaviour, that the views I examined claimed was necessary for agency, entailed that agents are not alienated from their actions, or that their actions conform to their deeply

[7] Here I am not committing to argue for such a cognitive penetration, but I rather wish to emphasize how a limited understanding of agentive awareness excludes some directions of research that could be fruitful for the theories examined.

[8] It is not necessary for the purposes of this work to develop a full-fledged account of the characteristics and different dimensions of self-consciousness. That is a huge topic and one that deserves attention in its own right, if these considerations about the role of self-consciousness for agency are convincing. In chapter 3 I mentioned Burge's view and the role he assigned to the 'I concept'. For alternative accounts see Bermudez, 2000, and Peacocke, 2014, among others.

held values, their practical identities etc. I argued in the last two chapters that this *augmented* sense of ownership is in fact a problem for these theories because of the kind of self-consciousness it entails and because it excludes immersed and mundane actions which do not display it. In a permissive view of agency such a conception of ownership does not play the same fundamental role. While it is an important phenomenon that we experience the distress of feeling alienated from our actions, the phenomenon of alienation does not put into question whether or not one is acting. Human beings who act cannot fail to be agents because their actions lack some capacity, or because they are estranged from their behaviour. Their actions can fail to possess certain qualities that we might find remarkable or desirable, and they are in this sense defective, but this does not entail that we question their status as actions: they might lack the characteristics that make them rational, autonomous, self-controlled, intentional etc. This means that, for example, in deeming a subject's behaviour as compulsive one is not denying that the subject is acting. Because these characteristics are not essential and constitutive features human agency, we can have an understanding of them that is independent of the definition of agency.

This is something that sharply distinguishes a permissive view of agency from the ones I have examined in this work. In fact, in defining one feature as that which grounds full agency, reflective endorsement views of the sort I have criticized conceive of several characteristics of actions as standing or falling together. So, for example, a self-governed agent is, on these views, one that is active, self-controlled and who identifies with his motives. And vice-versa, when someone is alienated from his motives, he is passive in a way that diminishes his agency. This sort of picture finds no support in a permissive view of agency because agency is not defined by self-governance, where this is conceived as encompassing these different aspects. The definition of actions thus holds no constitutive relation to one's behaviour being self-controlled, autonomous etc. I believe that this can be considered a positive upshot of a permissive view of agency because, in fact, it allows us to separate and distinguish elements that should be kept separate, and which the views I have criticized tend to identify. The result of this identification is that the

understanding of certain problematic cases remains obscure. To illustrate this, I will discuss briefly how different cases of behaviour that one could consider as not self-governed differ and are in fact complicated by the attempt to give a unified understanding of the different elements they present.

For example, an akratic person is not able to control himself and will act against his better judgement. Nevertheless, as we have seen in the last chapter, it is not always the case that the akratic person experiences alienation from his akratic behaviour. In fact, the akratic behaviour can be in harmony, rather than in conflict, with his self-image, despite the agent's regretting being swayed by these motives. In this case, he would not be alienated from his akratic actions. He might, regretfully, resignedly, accept that he just is the kind of person that behaves in such a way in the circumstances. Moreover, acting akratically does not exclude the possibility of one actively deciding to do something akratic.

On the other hand, the man who snaps at his friend and later claims that it was resentment speaking, not him, does not actively and deliberately decide to insult his interlocutor and he lacks self-control. Nevertheless, as stressed above, he might not lack self-understanding and he might recognize that he acted correspondingly to what he really thinks of the other person.

Moreover, consider someone so deeply committed to her family or religious values that she constantly suppresses all urges that go against them, some of which we could, as external observers, consider absolutely natural and healthy. A good example of this is the Victoria lady who experiences her normal sexual desires as alien and intrusive.[9] Every time this person faces a decision that sees her values and impulses clash, she successfully wins over her urges. This person does not display a lack of self-control, rather the opposite: she is overly controlled. She also genuinely identifies with these values and she actively decides to promote them. Nevertheless, we would feel that this person

[9] This is an example we find in Arpaly, 2003, 123. For a discussion of a similar case also see Velleman's analysis of Freud's Rat Man, (2006), 343-346.

is not in fact authentic or true to herself, and those urges she just casts aside are rather something she should face and accept.

The interaction between the notions of autonomy, authenticity, alienation and identification, and activity and passivity seem to vary in these cases. So, the akratic agent might fail at being self-controlled but is still active, and possibly authentic and not necessarily alienated from his motives. In the case of someone's angry reaction with his friend, we might see the person as passive but also authentic, despite his experiencing his reaction as something out of his control. In the last case I discussed, we can see that one can fail to be authentic, while still counting as self-governed, not alienated from one's motives and actively deciding upon them. In chapter 4 we also saw how anorexics, whom we don't consider as self-governed, might actually be self-controlled, not alienated from their motives, active in their deliberate decision to lose weight and authentic in their endorsing the value of lightness. A view that aims at connecting some of these elements to the constitutive conditions of agency needs to provide a clear picture of their interrelations and this can be, as we have seen in the above discussion of anorexia, a demanding task. A permissive view of agency need not deny the importance of these aspects or the interest in understanding their connections but it can in fact draw clearer distinctions among them because it does not strive to connect these different elements to the constitutive feature defining of agency altogether.

Moreover, a permissive view of agency can avoid an ambiguity that affects much of the debate around self-governance in the views we have examined, namely that it is not clear what it is for the agent to be "active", and in what way this can amount to being self-governed. On the one hand, on these views, activity and passivity seem to be related to the role played by reflection. We have seen how Korsgaard's conception of reflection in terms of deliberation creates a problem for her model, and we have also seen how Velleman struggles to account for the subject's employing a way of reflecting that is different from deliberation. The definition of activity and passivity has remained a point on which reflective endorsement models are ambiguous and I will briefly say why this is the case. In my sketch of the cases above I

employed the terms active and passive to characterize whether or not the subject deliberately decided for a certain course of action. I believe that within this understanding of activity we can in fact distinguish different dimensions: firstly, we can consider a subject as active because he actively engages in some form of practical reasoning; secondly, the subject can be active in deciding to undertake a certain action; and lastly, a subject can be active in executing his intention. A wanton would fail at being active because he would never actually decide and commit to some course of action, despite his being perfectly capable to reason practically. He would then be failing at the second stage I highlighted. An akratic, on the other hand, might perfectly well reason about what to do, and decide for the course of action he deems best, but then fail to actually carry out his intention with the result that he will do something against his better judgement. It is intuitive to think that a subject that is active in all three of these dimensions is more active than subjects that fail in at least one of them. This kind of view would hold these agents as more active in literally a quantitative sense. They are more active because they act more: they engage in the kind of activity that deliberation is, they take a decision and eventually execute what they have reasoned to do. In this sense, a subject who just passively acquires an intention, because a certain resolve dawns on him, maybe led by some emotion or just out of the blue, would be less active than this meticulously planning and deliberative agent. This idea of activity might seem straightforward but it is in fact problematic for the views I have examined.

Firstly, the connection between this kind of activity and self-control, for example, seems dubious. An agent who, through the steps just highlighted, resolves to follow a certain path of action is more active than one who passively acquires the intention to do that same thing. Is the former agent more self-controlled than the other? It is not clear that this is the case, because one can consider that someone who actively exercises self-control is a less self-controlled person than someone who simply does not need to. Even in terms of psychological unity, a factor that we have seen plays a crucial role in Korsgaard and Velleman, the agent who is not torn between alternatives and does not actively exercise self-control seems to enjoy more psychological unity

than one who does. The ambiguity of the relationship between self-control and the agent's activity is present in the theories examined and it springs from the obscure notion of control and the conception of reflection on which they rely. It seems to me that there is a fundamental difference between *being in control* and *exerting control* and it is not clear how the activity that seems entailed in the conception of reflective endorsement can take this distinction into account or have any place for it.

Secondly, what it is for an agent to be active is ambiguous in itself. The picture of activity just presented sees agents as more or less active in a rather quantitative manner: they are more active because they literally perform more activities. This definition does not correspond exactly to what we have seen Korsgaard and Velleman require in their conceiving agents as active when their actions are self-governed. Korsgaard sees agents as self-governed when they act by choosing their own ends in accordance with the categorical imperative. According to Velleman, self-governed agents act in accordance with reasons, which result from narrative coherence. It is true that these options are reserved for human behaviour: an animal's goals in its actions are not chosen as ends or reasons. But it is not clear why these agents should count as active. Sometimes our goals are set passively and unreflectively and even in the case in which agents display exemplary forms of activity and deliberation there are sets of biases, defaults and constraints that bound this activity, which thus itself also comprises passive elements that are not under my conscious control. As a result, the notion of activity is not clear and this bears on the definition of self-governance, since being active rather than passive is taken to be essentially associated with being self-governed.

Because a permissive view of agency need not hold a connection between these different elements, it gives the grounds for a discussion about their definitions and their relations that does not commit one to unify these different characteristics through its conception of agency.

Lastly, I will briefly consider a concern with such a permissive view and hint at why it is not in fact particularly problematic. Someone who holds a view

that does distinguish between kinds of agency and reserves a special role for a further sense of ownership of one's action can argue that a permissive view of agency is simply uninteresting because, for example, it is unclear that it can have any bearing on debates in ethics or moral responsibility. This concern emerges from the normative aspirations that, in my introduction, I have stressed reflective endorsement theories have. Since I have not addressed this topic, I have not discussed whether a conception of agency such as the one that results from reflective endorsement views can be fruitful when it comes to debates in moral philosophy. If, after examination, it turned out that reflective endorsement views of agency do not actually bear such results, the grounds of this objection would be weaker. This however would not provide an argument for the permissive view to defend itself from the charge. Nevertheless, one can question the very grounds on which the objection rests and argue that disentangling issues in the philosophy of action from those in moral philosophy would actually be a welcome result. None of these issues can be addressed here and they provide a direction for future work. However, I believe that it is important to stress how issues concerning self-governance, self-control, identification etc. need not cease being of great interest, just because one wishes to untie them from one's definition of agency. So, for example, undoubtedly the notion of identity remains central in ethics as well as in debates around multiculturalism,[10] even if one denies a constitutive relation between our practical identity and the conditions for agency. In fact, the kind of conception of practical identity that is defined by reflective endorsement views risks, on the one hand, to produce an unclear picture of identity by tying it to those elements that, I have claimed, would be best to keep separate; on the other hand, by privileging those aspects of identity which are connected to the agent's activity, deliberation and self-control, this conception does not provide a full understanding of the biases and constraints that are at play in defining one's identity, nor does it make sense of the role that identity can play in automatic behaviour.[11]

[10] See Sen, 2006.

[11] So, for example, aspects concerning one's identity figure in strategies such as MINDSPACE (Dolan, 2012) developed in behavioural economics. These kinds of strategies rely on the idea that one can integrate defaults and biases within one's model and methods and employ factors that operate on the automatic systems. This kind of method relies on contextual changes to

Similarly, a permissive view of agency need not deny the importance of narratives in our lives. Indeed, even without making narrativity a condition for agency, there are several reasons why the role of narratives should not be belittled. Our narrative practices are at the core of our thinking about our past and future and play an important role in understanding ourselves and others because they organize and integrate our experiences, enabling us to make sense of ourselves, our actions and the actions of others. It is of great interest how the role of narratives can be essential in certain contexts,[12] and more generally in living flourishing lives as individuals.

bring about behaviour change, rather than striving to change the subjects' beliefs and attitudes.

[12] For example, in the context of understanding and treating certain illnesses, such as schizophrenia, see McKenzie and Poltera, 2010.

Bibliography

ALLISON, H. (1983), *Kant's Transcendental Idealism,* (Connecticut: Yale University Press).

—. (2004) *Idealism and Freedom. Essays on Kant's Theoretical and Practical Philosophy,* (Cambridge: CUP).

ARONSON, E. (1969). The theory of cognitive dissonance: A current perspective, in *Advances in experimental social psychology,* L. Berkowitz (Ed.), (New York: Academic Press), 4: 1-34.

ARPALY, N. (2003), *Unprincipled Virtue,* (Oxford: OUP).

ARPALY, N. AND SCHROEDER, T. (1999), 'Alienation and Externality' in *Canadian Journal of Philosophy*, 29 (3): 371-387.

—. (2012), 'Deliberation and Acting for Reasons' in *Philosophical Review,* 121(2): 209-239.

AUSTEN, J. (2009), *Emma*, (New York: Wild Jot Press).

BAUMEISTER, R.F. (2002), 'Ego Depletion and Self-Control Failure: An Energy Model of the Self's Executive Function', *Self and Identity*, 1: 129–136.

BAUMEISTER, ROY F. AND E.J. MASICAMPO (2010), 'Conscious thought is for facilitating social and cultural interactions: How simulations serve the animal-culture interface', *Psychological Review* 117 (3): 945-71.

BAYNE, T. (2006), 'Phenomenology and the feeling of doing: Wegner on the conscious will', in *Does Consciousness Cause Behaviour?,* S. Pockett, W.P. Banks, and S. Gallagher (eds.), (MA:MIT), reprinted (2009), pp. 169-186.

—. (2011), 'The Sense of Agency', in *The Senses: Classic and Contemporary Philosophical Perspectives*, Fiona Macpherson (ed.), (Oxford: OUP), 355-374.

BAYNE, T. AND LEVY, N.,(2005), 'Amputees By Choice: Body Integrity Identity Disorder and the Ethics of Amputation', *Journal of Applied Philosophy,* 22 (1): 75–86.

BAYNE, T. AND PACHERIE, E. (2007), 'Narrators and Comparators: the architecture of agentive awareness', *Synthese*, 159:475-491. doi: 10.1007/s11229-007-9239-9

BAYNE, T., CLEEREMANS, A. AND WILKEN, P. (2009*), The Oxford Companion to Consciousness*, (Oxford: OUP).

BEM, D.J. (1972), 'Self Perception Theory', in *Advances in Experimental Social Psychology,* L. Berkowitz (ed)., (New York: Acedemic Press), 6: 1-62.

BERMUDEZ, J. (2000), *The Paradox of Self-Consciousness*, (MA: MIT Press).

BRATMAN, M. (2007), *Structures of Agency: Essays*, (Oxford: Oxford University Press).

BURGE, T. (2009), 'Primitive agency and natural norms', *Philosophy and Phenomenological Research*, 79 (2), 251-278.

CARRUTHERS, P. (2007). 'The illusion of conscious will', *Synthese*. doi: 10.1007/s11229-007-9204-7.

—. (2009). 'How we know our own minds: The relationship between mindreading and metacognition' in *Behavioral and Brain Sciences*, 32:121-138.

COHEN, G.A., (1996), 'Reason, Humanity and the Moral Law', in *The Sources of Normativity,* (Cambridge: CUP). 167-170.

DAVIDSON, D., (1987), 'Knowing One's Own Mind' in *Proceedings and Addresses of the American Philosophical Association,* 61: 441–58.

—. (1985), 'Deception and Division', in *Actions and Events*, E. LePore and B. McLaughlin (eds.), (New York: Basil Blackwell).

DE NEYS, W. AND GLUMICIC, T. (2008), 'Conflict monitoring in dual process theories of thinking', *Cognition,* 106, 1248–1299.

DE RIDDER ET AL. (2008), 'Psychological adjustment to chronic disease', *The Lancet*, 372, 246-255.

DENNETT, D. (1991), *Consciousness Explained*, (London: Penguin Books).

DOLAN, P. (2012), 'Influencing Financial Behavior: From Changing Minds to Changing Contexts', *Journal of Behavioral Finance*, 13:2, 126-142, doi: 10.1080/15427560.2012.680995

DREYFUS, H.L, (2002), 'Refocusing the question: can there be skillful coping without propositional representations or brain representations?', *Phenomenology and the Cognitive Sciences*, 1: 413-425.

ENOCH, D., (2006), 'Agency, Shmagency: Why Normativity Won't Come from What is Constitutive of Agency', *Philosophical Review*, 115: 169–198.

EVANS, J. ST. B. T AND STANOVICH, K., (2013), 'Dual-Process Theories of Higher Cognition: Advancing the Debate', *Perspectives on Psychological Science,* 8(3): 223–241. doi: 10.1177/1745691612460685

EVANS, J. ST. B. T. (1984), 'Heuristic and analytic processing in reasoning'.

British Journal of Psychology, 75, 451–468.

—. (2008), 'Dual-processing accounts of reasoning, judgment and social cognition', *Annual Review of Psychology*, 59, 255–278.

FESTINGER, L. AND CARLSMITH, J.M. (1959), 'Cognitive Consequences of Forced Compliance', *Journal of Abnormal and Social Psychology, 58*:203-210.

FOURNERET, P. AND JEANNEROD, M. (1998), 'Limited conscious monitoring of motor performance in normal subjects', *Neuropsychologia*, 36, 1133–1140.

FRANKFURT, H. (1971), 'Freedom of the Will and the Concept of a Person', in *The Importance of What we Care About,* Cambridge: Cambridge University Press, reprinted (2009).

—. (1998), 'The importance of what we care about', in *The Importance of What We Care About: Philosophical Essays,* (Cambridge: CUP), 80-94. Reprinted (2009).

FRIEDERICH, H.C, AND HERZOG, W. (2011), 'Cognitive-Behavioral Flexibility in Anorexia Nervosa', *Current Topics in Behavioral Neurosciences,* 6: 111-123.

GALLAGHER, S. (2000), 'Philosophical conceptions of the self: implications for cognitive science', *Trends in Cognitive Science,* 4 (1): 14-21.

—. (2003), 'Self-narrative, Embodied Action, and Social Context', in *Between Suspicion and Sympathy: Paul Ricoeur's Unstable Equilibrium (Festschrift for Paul Ricoeur),* A. Wiercinski (ed.), (Toronto: The Hermeneutic Press), 409-423.

—. (2012), 'Multiple aspects of agency', *New Ideas in Psychology* 30: 15-31. doi:10.1016/j.newideapsych.2010.03.003

GALLAGHER, S. AND MARCEL, A. (1999), 'The Self in Contextualized Action' in *Models of the Self,* S. Gallagher and J. Shear (eds.), (Exeter: Imprint Academic), pp. 273-300. Reprinted (2002).

GARYFALLOS, G. ET AL. (2010), 'Comorbidity of obsessive–compulsive disorder with obsessive–compulsive personality disorder: Does it imply a specific subtype of obsessive–compulsive disorder?',*Psychiatry Research,* 177 (1–2): 156–160.

GAZZANIGA M.AND LEDOUX, J.(1978), *The Integrated Mind*, (New York: Plenum Press).

GAZZANIGA, M. (1985), *The Social Brain: Discovering the Networks of the Mind*, (New York: Basic Books).

GIORDANO, S., (2005), *Understanding Eating Disorders: Conceptual and EthicalIssues in the Treatment of Anorexia and Bulimia Nervosa*, (Oxford: Clarendon Press).

GOLDIE, P (2012), *The Mess Inside: Narrative, Emotion and Mind*, (Oxford: OUP).

GOLDMAN, A. (1993), 'The Psychology of Folk-Psychology', *Behavioural and Brain Sciences*, 16: 15-28.

GRAHAM, G., AND STEPHENS, G. L. (1994), 'Mind and mine', in, *Philosophical psychopathology*, G. Graham, & G. L. Stephens (Eds.), (MA: MIT Press), 91–109.

HAGGARD, P. (2005), 'Conscious intention and motor cognition', *Trends in Cognitive Sciences,* 9(6): 290–295.

HARRÉ, R. (1984), *Personal Being: A theory for transcendental psychology,* (Cambridge: Harvard University Press).

HIRSTEIN, W. (2006), Brain Fiction, Self-deception and the Riddle of Confabulation, (MA: MIT Press).

HOLTON, R. (2009), *Willing, Wanting, Waiting*, (Oxford:OUP).

HORNSBY, J. (2004),'Agency and actions', *Royal Institute of Philosophy Supplement* 55, 1-23.

HUTTO, D. (2007), *Narrative and Understanding Persons,* (Cambridge: CUP).

JAMES, W. (1890), The Principles of Psychology, Vol.1, (New York: Cosimo Inc.). Reprinted (2007).

JEANNEROD, M. (1997), *The cognitive neuroscience of action,* (Oxford: Blackwell).

JOSEPH, R. (1999), 'Frontal lobe psychopathology: mania, depression, confabulation, catatonia, perseveration, obsessive compulsions, and schizophrenia', *Psychiatry*. 62(2): 138-72.

KAHNEMAN, D., & FREDERICK, S. (2002), 'Representativeness revisited: Attribute substitution in intuitive judgement' in *Heuristics and biases: The psychology of intuitive judgement*, T. Gilovich, D. Griffin, & D. Kahneman (Eds.), (Cambridge: CUP), 49–81.

KANT, I. (1956), *Critique of Practical Reason,* trans. L.W. Beck (Indianapolis and New York: Bobbs-Merrill).

—. (1960), *Religion within the Limits of Reason Alone,* trans. T.M. Greene and H. H. Hudson (New York: Harper and Row).

—. (1965), *Critique of Pure Reason,* trans. N. Kemp Smith (New York: St. Martin's Press).

—. (1991), *The Metaphysics of Morals*, trans. M. Gregor, (Cambridge: CUP).

KENNETT, J. (2013), 'Just Say No?', in *Addiction and Self-Control: Perspectives from Philosophy, Psychology and Neuroscience*, N. Levy (ed.), (Oxford: OUP), 144-164.

KENNETT, J. AND FINE, C. (2009), 'Will the Real Moral Judgment Please Stand up? The Implications of Social Intuitionist Models of Cognition for Meta-Ethics and Moral Psychology', *Ethical Theory and Moral Practice*, 12 (1): 77-96.

KORSGAARD, C. (1989), 'Personal Identity and the Unity of Agency: A Kantian Response to Parfit', *Philosophy & Public Affairs*, 18(2):101-132.

—. (1996), The Sources of Normativity, (Cambridge: CUP).

—. (1999), "Self-constitution in the ethics of Plato and Kant" in *Journal of Ethics,* 3 (1): 1-29.

—. (2008), *The Constitution of Agency*, (Oxford: OUP).

—. (2009*), Self-constitution. Agency, Identity and Integrity,* (Oxford: OUP).

—. (FORTHCOMING), 'The normative constitution of agency' in *Rational and Social Agency: Essays on the Philosophy of Michael Bratman,* M. Vargas and G. Yaffe (eds.), (New York: OUP).

LAU, H. C., ROGERS, R. D., HAGGARD, P., AND PASSINGHAM, R.E. (2004), 'Attention to intention', *Science*, 303: 1208–1210.

LECKY, P. (1945*), Self-Consistency: A Theory of Personality,* (New York: Island Press).

LEIBNIZ, G.W. (1989), *Philosophical Essays*, trans. R. Ariew and D. Garber (eds.), (Indianapolis and Cambridge: Hackett Publishing Company).

LEVY, N. (2006), 'Addiction, autonomy and ego-depletion: A response to Bennett Foddy and Julian Savulescu', *Bioethics,* 20 (1): 16–20.

—. (2011), 'Resisting 'Weakness of the Will'', *Philosophy and Phenomenological Research,* 82 (1): 134-155.

—. (2013a), 'The Importance of Awareness', *Australasian Journal of Philosophy,* 91 (2), 211-229. doi: 10.1080/00048402.2012.684883

—. (2013b) 'Addiction and Self-Control: Perspectives from Philosophy,

Psychology and Neuroscience' in *Addiction and Self-Control: Perspectives from Philosophy, Psychology and Neuroscience*, N. Levy (ed.), (Oxford: OUP), 1-15.

LEVY, N. AND BAYNE, T. (2004), 'Doing without deliberation: automatism, automaticity, and moral accountability', *International Review of Psychiatry,* 16(3), 209–215.

LHERMITTE F. (1983), 'Utilization behavior and its relation to lesions of the frontal lobes', *Brain,* 106: 237-255.

MACINTYRE, A. (2007), *After Virtue,* (Notre Dame: University of Notre Dame Press).

MACKENZIE, C. (2007), 'Bare Personhood? Velleman on selfhood' **in** *Philosophical Explorations,* 10 (3): 263-282.

MACKENZIE, C.. AND POLTERA, J. (2010), 'Narrative Integration, Fragmented Selves, and Autonomy', *Hypatia*, 25 (1): 31-54.

MARCEL, A. (2003), 'The Sense of Agency: Awerness and Ownership of Action' in *Agency and Self-Awareness*, J. Roessler and N. Eilan (eds.), (Oxford: OUP), 48-93. Reprinted (2008).

MARCHETTI, C. AND DELLA SALA, S. (1998), 'Disentangling the Alien and Anarchic Hand', Cognitive Neuropsychiatry, 3 (3): 191-207.

MCDOWELL, J. (1996), *Mind and World,* (MA: Harvard University Press).

MELE, A. (2001), *Self-Deception Unmasked,* (Princeton: Princeton University Press).

MILLER, R.L., BRICKMAN, P. AND BOLEN, D. (1975), 'Attribution Versus Persuasion as a Means for Modifying Behaviour', *Journal of Personality and Social Psychology*, 31(3): 430-441.

MONSELL, S. (1996), 'Control of Mental Processes', in *Unsolved Mysteries of the Mind: Tutorial Essays in Cognition*, V. Bruce (ed.), (Hove, UK: Lawrence Erlbaum Associates Ltd.), 93-148.

NEIMEYER, R.A. (2003), 'Community and Coherence: Narrative Contributions to the Psychology of Conflict and Loss' in *Narrative and Consciousness: Literature, Psychology and the Brain*, G.D. Fireman, T.E. McVay, and O.J. Flanagan (eds.) (Oxford: OUP).

NISBETT, R.E. AND VALINS, S. (1972), 'Perceiving the causes of one's own behaviour' in *Attribution: Perceiving the Causes of Behaviour*, E.E. Jones *et al.* (eds.), (Morristown: General Learning).

NISBETT, R.E. AND WILSON, T.D. (1977), 'Telling more than we can know: Verbal reports of mental processes', *Psychological Review*, 84: 231-259.

O'BRIEN K.M., AND VINCENT N.K,. (2003), 'Psychiatric comorbidity in anorexia and bulimia nervosa: nature, prevalence, and causal relationships', *Clinical Psychology Review*, 23(1): 57-74.

OLSON, E. (1997), *The Human Animal: Personal Identity Without Psychology*, (Oxford: OUP).

OSMAN, M. (2004), 'An evaluation of dual-process theories of reasoning', *Psychonomic Bulletin & Review,* 11 (6): 988-1010.

OWENS, D. (2011), 'Deliberation and the first person'. in *Self-Knowledge*, A. Hatzimoysis, (ed.), (Oxford University Press, Oxford), pp. 261-277.

PACHERIE, E. (2007a), 'The Sense of Control and the Sense of Agency', *Psyche* 13 (1): 1-30.

—. (2007b), 'The anarchic hand syndrome and utilization behavior: a window onto agentive self-awareness', *Functional Neurology*, 22 (4): 211-217.

—. (2010), 'Self-Agency' in *The Oxford Handbook of the Self*, S. Gallagher (Ed.), (Oxford: OUP), 440-462.

—. (2011), 'Nonconceptual representations for action and the limits of intentional control', *Social Psychology*. 42(1): 67-73.

—.. (2012), 'Action', in *The Cambridge Handbook of Cognitive Science*, Keith Frankish and William Ramsey (eds.), (Cambridge: CUP), 92-111.

PARFIT, D. (1984), *Reasons and Persons*, (Oxford: OUP).

PEACOCKE, C. (1998), 'Conscious Attitudes, Attention and Self-Knowledge' in C. Wright, B. Smith and C. MacDonald (eds.) *Knowing Our Own Minds* (Oxford: OUP).

—. (2014*), The Mirror of the World*, (Oxford: OUP).

PICKARD, H. AND PEARCE, S. (2013), 'Addiction in Context' in *Addiction and Self-Control: Perspectives from Philosophy, Psychology and Neuroscience*, N. Levy (ed.), (Oxford: OUP), 165-189.

PUJOL, J. ET AL. (2004), 'Mapping structural brain alterations in obsessive-compulsive disorder', *Archives of General Psychiatry,* 61(7): 720-30.

REASON, J. (1984), 'Lapses of attention in everyday life', in *Varieties of Attention*, R. Parasuraman and D.R. Davies (eds.), (Orlando: Academic Press), 515-550.

RICOEUR, P. (1988), *Time and Narrative,* trans. K. McLoughlin and D. Pellauer, (Chicago: University of Chicago Press).

—. (1992), *Oneself as Another*, trans. Kathleen Blamey, (Chicago: University of Chicago).

ROESSLER, J. (2013), 'The Epistemic Role of Intentions', *Proceedings of the Aristotelian Society*,113 (1): 41-56.

ROSER, M.E. AND GAZZANIGA, M.S. (2006), 'The interpreter in Human Psychology' in *Evolution of Nervous Systems, Volume V: The Evolution of Primate Nervous Systems*, T. M. Preuss & J.H. Kaas (eds.), Academic Press: Oxford.

SAINSBURY, R.M. (2010), *Fiction and Fictionalism,* (New York: Routledge).

SASS, L. (1992). *Madness and modernism: Insanity in the light of modern art, literature, and thought*. New York: Basic Books.

SCHECHTMAN, (2007) 'Stories, Lives, and Basic Survival: A Refinement and Defense of the Narrative View' in *Narrative and Understanding Persons,* D. Hutto (ed.), (Cambridge: CUP), 155-178.

—. (1996), *The Constitution of Selves,* (New York: Cornell University Press).

SEN, A. (2006), *Identity and Violence: the illusion of destiny,* (London: Penguin Books).

SHINDLER, A.G, et al. (1984), 'Intrusions and Perseverations', *Brain and Language.,* 23(1): 148-58.

SHOEMAKER, S. (1984), 'Personal Identity: A Materialist's Account', in S. Shoemaker and R.Swinburne, *Personal Identity*, (Oxford: Blackwell).

—. (1996), *The First-Person Perspective and Other Essays* (Cambridge: CUP).

SLOMAN, S.A. (1996), 'The empirical case for two systems of reasoning', *Psychological Bulletin,* 119, 3–22.

SNOWDON, P. (1995) 'Persons, animals and bodies' in *The Body and the Self*, in J.L. Bermudez, A.J. Marcel & N.M. Eilan (eds.), (MA: Mit Press).

STANOVICH, K. E. (1999), *Who is rational? Studies of individual differences in reasoning,* (Mahwah, NJ: Erlbaum).

STEPHENS, G. L., & GRAHAM, G. (2000). *When self-consciousness breaks: Alien voices and inserted thoughts*, (Cambridge: MIT Press).

STRAWSON, G. (2002), 'The Self', in *Models of the Self,* S. Gallagher and J. Shear (eds.), (Exeter: Imprint Academic), pp. 1-24.

STRAWSON, G. (2004) 'Against narrativity'. *Ratio*, 17 (4): 428-52.

—. (2012) 'We live beyond any tale that we happen to enact'. *Harvard Review of Philosophy,* 18:73-90.

SWANN , W.B (1986), 'To be adored or to be known? The interplay of self-enhancement and identity disruption', *Handbook of Motivation and Cognition*, R.M. Sorrentino and E.T. Higgins (eds.), (New York: Guilford Press), 408-448.

TSAKIRIS, M., AND HAGGARD, P. (2005), 'Experimenting with the acting self', *Cognitive Neuropsychology,* 22: 387–407.

TSAKIRIS, M., SCHUTZ-BOSBACH, S. AND GALLAGHER, S. (2007), 'On agency and body-ownership: Phenomenological and neurocognitive reflections', *Consciousness and Cognition*, 16: 645–660.

VALLACHER, R.R. AND WEGNER, D.M (1985), *A theory of action identification,* (New York: Lawrence Erlbaum Associates).

VAN IMWAGEN, P. (2003), 'Existence, Ontological Commitment and Fictional Entities' in *The Oxford Handbook of Metaphysics*, M.J. Loux and D.W. Zimmerman, (Oxford: OUP), 131-157.

VELLEMAN, D. (2000), *The Possibility of Practical Reason*, (University of Michigan: OUP).

—. (2003), 'Narrative Explanation' in *The Philosophical Review,* 12 (1): 1-25.

—. (2004a), 'Précis of "The Possibility of Practical Reason"' in *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition*, 121 (3): 225-238.

—. (2004 b), 'Replies to Discussion on "The Possibility of Practical Reason' in *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition*, 121 (3): 277-298.

—. (2006), *Self to Self: Selected Essays*, (Cambridge: CUP).

—. (2007a), *Practical Reflection*, (Stanford: CSLI Publications).

—. (2007b), 'Reply to Catriona Mackenzie', *Philosophical Explorations: An International Journal for the Philosophy of Mind and Action,* 10 (3): 283-290.

—. (2008), 'Bodies, Selves', *American Imago*, 65 (3): 405-426.

—. (2013), *Foundations for Moral Relativism, (*OpenBook Publishers).

WEGNER, D. M. (2002). *The Illusion of Conscious Will*. Cambridge, (MA: MIT Press).

WEGNER, D.M. AND VALLACHER, R.R. (1986), 'Action Identification' in *Handbook of Motivation and Cognition*, R.M. Sorrentino and E.T. Higgins (eds.), (New York: Guilford Press), 550-582.

WHEELER, S.C., ET AL. (2007), 'Resistance to per- suasion as self-regulation: Ego depletion and its effects on attitude change processes', *Journal Of Experimental Social Psychology,* 43: 150–156.

WRIGHT, C. (1998), 'Self-knowledge: the Wittgenstenian Legacy' in C. Wright, B. Smith and C. MacDonald (eds.) *Knowing Our Own Minds* (Oxford: OUP).

YARYURA-TOBIAS, J.A. AND NEZIROGLU, F. (1983*), Obsessive Compulsive Disorders: Pathogenesis, Diagnosis and Treatment*, (New York: Marcel Dekker).

ZAHAVI,D. (2007) 'Self and Other.The limits of narrative understanding' in Hutto, D. (2007), *Narrative and Understanding Persons*, (Cambridge: CUP), 179-202.

ZALTA, E.N. (1983), *Abstract Objects: An Introduction to Axiomatic Metaphysics*, (Dordrecht: Reidel).